RICE UNIVERSITY

High-dimensional and dependent data with additional structure

by

Sergii Babkin

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Michael Schweinberger, Chair
Assistant Professor of Statistics

Katherine B. Ensor
Professor of Statistics

Richard A. Tapia
University Professor
Maxfield-Oshman Professor of Engineering
Professor of Computational and Applied Mathematics

Houston, Texas

April, 2017

ABSTRACT

High-dimensional and dependent data with additional structure

by

Sergii Babkin

The age of computing has enabled the collection of massive amounts of data. These data present numerous statistical challenges, because many data sets are high-dimensional and dependent. While statistical inference for high-dimensional and dependent data is challenging, many data come with additional structure that can be exploited to facilitate statistical inference. This thesis considers two widely used classes of models for high-dimensional and dependent data with additional structure, high-dimensional multivariate time series and exponential-family random graph models.

In the case of high-dimensional multivariate time series, there is often additional structure in the form of spatial structure, e.g., air pollution is monitored by monitors and the geographical locations of monitors are known. If air pollutants cannot travel long distances, then the estimation of past-present and present-present dependencies of air pollution at monitors can be restricted to short distances. Here, a novel two-step estimation approach is proposed to estimate the range of dependence along with the parameters of multivariate time series in high-dimensional settings. Theoretical results show that the two-step estimation approach reduces statistical error in high-dimensional settings. Simulation results confirm that the two-step estimation approach reduces statistical error and computing time. An application to air pollution in the U.S. demonstrates that the two-step estimation approach gives rise to results that are in line with scientific knowledge,

whereas estimation approaches ignoring the spatial structure report results that are in conflict with scientific knowledge.

In the case of exponential-family random graph models, it is likewise common that there is additional structure: e.g., it is known that many networks, such as insurgencies and terrorist networks, are local in nature. Here, a novel two-step estimation approach is proposed to estimate the local structure along with the dependence pattern of networks. The proposed two-step estimation approach can be implemented in parallel and hence paves the ground for massive-scale estimation of exponential-family random graph models. Theoretical results are provided along with simulation results. An application to a large Amazon product network demonstrates the usefulness of the proposed two-step estimation approach.

Keywords: Dependent data; High-dimensional data; Vector autoregressive process; Exponential-family random graph model; Local dependence.

Acknowledgments

I would like to thank my advisor, Dr. Michael Schweinberger for his expertise, direction, and assistance throughout my Ph.D. program. Without his constant support and motivation, this work would not have been possible. In addition, most of my research was supported by Dr. Schweinberger's NSF award DMS-1513644.

I would also like to thank the other members of my committee, Dr. Katherine B. Ensor and Dr. Richard A. Tapia, for their support throughout my Ph.D. program. Their commitment to my success was invaluable. Dr. Katherine B. Ensor served as an excellent mentor, always willing to provide guidance and advice.

Thanks is also due to my student colleagues and friends who were there for me both in courses as well as in life. Special mention goes to Oleg Melnikov, Daniel Cross, Jonathan Stewart, and Robert Kosar for their help in our classes and insightful discussions.

Finally, and most importantly, I would like to thank my wife Daria. Her never-failing support, encouragement, quiet patience, and unwavering love were undeniably the bedrock upon which the past eight years of my life have been built. Also, I would like to thank my family for their faith in me and allowing me to be as ambitious as I wanted. I would not be where I am today without them.

Contents

Illustrations

Tables

Chapter 1

Introduction

This thesis develops large-scale statistical methods for two classes of models for high-dimensional and dependent data with additional structure, vector autoregressive processes (VAR) and exponential-family random graph models (ERGMs). VAR processes were popularized by Sims (1980) and are widely used for studying the complex interrelationships among the components of multivariate time series (Lütkepohl, 2011), such as the understanding of the human brain (Friston, 2009). ERGMs were pioneered by Holland and Leinhardt (1981) and Frank and Strauss (1986) and are widely used by network scientists for modeling complex dependence in network data (Lusher et al., 2013), such as network redundancy and vulnerability in insurgencies and terrorist networks (Koschade, 2006).

While statistical inference for such high-dimensional and dependent data is challenging, many data come with additional structure that can be exploited to facilitate statistical inference. In the case of VAR processes, there may be additional structure in the form of spatial structure. An example is given by air pollution monitored at hundreds of locations across the U.S.A. It is well-known that air pollution at one location may affect air pollution at neighboring locations, but not at distant locations. That suggests that statistical inference can be facilitated by confining the estimation of dependencies between the components of VAR processes to short distances. If the range of dependence is unknown, it has to be estimated. Chapter 2 proposes large-scale statistical methods for doing so. In the case of ERGMs, it is likewise plausible that there is additional structure: e.g., it is well-known that many networks, such as insurgencies and terrorist networks, are local in nature. If networks are local in nature, the estimation of dependencies can be confined to subnetworks. If the local structure is unknown, it can be estimated. Chapter 3 elaborates the first large-scale statistical methods for doing so.

The remainder of the thesis is structured as follows. The remainder of Chapter 1 reviews VAR processes and ERGMs and discusses the contributions of this thesis to the study of these models. Chapters 2 and 3 discuss large-scale statistical methods for VAR processes and ERGMs with additional structure, respectively. Chapter 4 discusses open problems and gives directions for future research. All data and source code used in this thesis are publicly available, as described in Chapter 5.

1.1 High-dimensional multivariate time series with additional structure

Chapter 1 is concerned with estimating high-dimensional multivariate time series using VAR processes. Let $\boldsymbol{X}(t) = (X_1(t), \ldots, X_k(t))_{t=1}^N$ denote a k-dimensional L-th order VAR process of the form

$$\boldsymbol{X}(t) \;=\; \sum_{l=1}^{L} \boldsymbol{A}_l \, \boldsymbol{X}(t-l) + \boldsymbol{e}(t),$$

where $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L$ are $k \times k$ transition matrices and the errors $\boldsymbol{e}(t)$ are independent multivariate Gaussian random variables with mean $\mathbf{0}_k$ and positive-definite error covariance matrix $\boldsymbol{\Sigma}$. Transition matrices capture temporal (i.e., past-present) relationships among the individual system components, while the error covariance matrix captures additional contemporaneous (i.e., present-present) dependencies among them. This thesis concentrates on the problem of estimation of the transition matrices, since they are essential for structural analysis and simultaneous forecasting of the components of a multivariate time series (e.g., Stock and Watson, 2006; Bańbura et al., 2010). This problem amounts to recovering $p = k^2 L$ parameters corresponding to all possible pairs of components of the multivariate times series for all possible values of time lag.

Much work has been done on VAR processes in classical low-dimensional settings (e.g., Watson, 1994; Waggoner and Zha, 1999; Brillinger, 2001; Lütkepohl, 2005, 2011). In high-dimensional settings, some early consistency results on models with ℓ_1- and other penalties were obtained by Song and Bickel (2011) and Negahban and Wainwright (2011), though both made strong assumptions. The most notable contributions were made by Loh and Wainwright (2012) and Basu and Michailidis (2015) who developed powerful concentration inequalities for a new class of M-

estimators. They established consistency under weak conditions and showed that these conditions are satisfied with high probability. In particular, Basu and Michailidis (2015) derived non-asymptotic upper bounds on the estimation errors for ℓ_1-penalized least squares estimators under high-dimensional scaling. Bayesian researchers have used priors to endow high-dimensional VAR process with low-dimensional structure (De Mol et al., 2008; Koop, 2013). Davis et al. (2016) proposed a likelihood-based approach and a two-stage estimation procedure encouraging sparsity.

The mentioned theoretical results assume that multivariate time series do not have additional structure. However, many multivariate time series do have additional structure in practice: e.g., air pollution monitors have locations. Likewise, the human brain is structured and brain cells have positions in the human brain. Such structure can be exploited to reduce computing time and statistical error. In contrast, if such structure is ignored, high-dimensional methods can produce results that are not meaningful from a scientific point of view. Chapter 2 gives an example to demonstrate that.

In many applications, dependence in a multivariate time series is local in the sense that distances between dependent components of the multivariate time series are substantially shorter than the largest distance between any pair of components. Thus, if maximum distance between dependent components $d_{\max}$ is known, model estimation can be restricted to only pairs of components separated by distances $d \leq d_{\max}$. Hence, the following two-step estimation approach is proposed. First step estimates $d_{\max}$ if it is unknown, while the second step utilizes the estimated $d_{\max}$ to estimate local dependencies among the components of the multivariate time series. This approach greatly reduces computing time and the statistical error of the parameters given that $d_{\max}$ is sufficiently short, the components are not too close to each other, and the dimensionality of the multivariate time series is sufficiently large. An important feature of the proposed procedure is that no assumptions are made about the parametric form of relationship between the distance separating the components of the VAR process and the strength of dependence between the components. Additionally, this thesis provides non-asymptotic error bounds on the estimates of the transition matrix parameters that hold with high probability and show that the proposed two-step approach

reduces the statistical error.

1.2 Exponential-family random graph models with additional structure

Chapter 3 is concerned with with exponential-family random graph models. To describe ERGMs, let $\mathcal{A}$ be a set of nodes and $X_{i,j} = 1$ if there is an edge between nodes i and j and $X_{i,j} = 0$ otherwise. Denote by $\boldsymbol{X} = (X_{i,j})$ and by $\mathbb{X}$ the set of possible values of $\boldsymbol{X}$. ERGMs assume that a random graph is governed by an exponential family (Brown, 1986) with support $\mathbb{X}$ and probability mass functions of the form

$$p_{\boldsymbol{\eta}}(\boldsymbol{x}) \;\; = \;\; \exp(\langle \boldsymbol{\eta},\, s(\boldsymbol{x}) \rangle - \psi(\boldsymbol{\eta})), \qquad \boldsymbol{x} \in \mathbb{X},$$

where $\langle \boldsymbol{\eta},\, s(\boldsymbol{x}) \rangle$ is the inner product of a natural parameter vector $\boldsymbol{\eta} \in \mathbb{R}^{\dim(\boldsymbol{\eta})}$ and a vector of sufficient statistics $s : \mathbb{X} \mapsto \mathbb{R}^{\dim(\boldsymbol{\eta})}$ and $\psi(\boldsymbol{\eta})$ is the log-normalizing constant. ERGMs are flexible models that admit a wide range of dependence structures through the choice of the sufficient statistic vector $s(\boldsymbol{x})$ (Lusher et al., 2013). Examples of sufficient statistics include counts of edges $x_{i,j}$, two-stars $x_{i,j}\, x_{i,k}$, triangles $x_{i,j}\, x_{j,k}\, x_{i,k}$, and various other functions of a graph. In practice, sufficient statistics that include interactions of edge variables are of great interest, since they induce dependence among edges. In addition, sufficient statistics incorporating various node attributes are often used (e.g., gender, race, wealth, etc.).

However, while ERGMs can be most useful, it has turned out in the past decade that some ERGMs have important problems. One of the problems is that as the size of the network grows, some ERGMs can induce strong long-range dependence (Strauss, 1986; Jonasson, 1999; Häggström and Jonasson, 1999; Handcock, 2003). This can result in model degeneracy (Handcock, 2003; Rinaldo et al., 2009), the property of models to place most of the probability mass on extreme graphs, such as empty or complete graphs (Schweinberger, 2011; Chatterjee and Diaconis, 2013). Graphs generated from such degenerate models tend to be close to the boundary of the convex hull of the sufficient statistic vector. This can lead to problems in maximum likelihood estimation, because maximum likelihood estimators do not exist when the sufficient statistic vector is on the boundary

of the convex hull (Handcock, 2003; Rinaldo et al., 2009). Even when maximum likelihood estimators exist, finding them by numerical methods may be challenging when the observed sufficient statistic vector is close to the boundary of the convex hull (Handcock, 2003; Rinaldo et al., 2009). In addition, Shalizi and Rinaldo (2013) suggested that some classes of exponential-family random graph models are not projectable in the sense that the marginal distributions of subgraphs under those models may not be consistent with the distribution of the whole graph. As a result, statistical inference for exponential-family random graph models may not be sensible.

To address these issues, Schweinberger and Handcock (2015), Schweinberger and Stewart (2016), and Schweinberger (2017) suggested to endow exponential-family random graph models with additional structure. The basic idea is that networks are local in nature (e.g., Granovetter, 1973; Wasserman and Faust, 1994; Pattison and Robins, 2002) and hence it makes sense to assume that a large graph is composed of independent subgraphs with local dependence. For example, insurgencies consist of groups of local fighters and likewise terrorist networks consist of groups of terrorists. The relationships within one group of local fighters may be dependent, but are unlikely to depend on relationships to local fighters belonging to other groups.

Endowing exponential-family random graph models with additional structure has multiple advantages including local depende. First, local dependence induces weak dependence and models with weak dependence are less prone to model degeneracy. Schweinberger and Stewart (2016, Corollary 1) showed that under weak conditions models with local dependence place most probability mass on graphs that are far from empty or complete graphs and hence are not prone to model degeneracy. Second, models with local dependence satisfy a weak form of self-consistency in the sense that the probability mass function of a subgraph is consistent with the probability mass function of the whole graph (Schweinberger and Handcock, 2015, Theorem 1). This property enables consistent estimation of neighborhood-dependent parameters. Schweinberger and Stewart (2016) also showed that M-estimators of both canonical and curved exponential-family random graph models with local dependence are consistent when the neighborhoods are observed and grow at the same rate. An important practical problem is that in most networks the neighborhood structure

is unobserved and has to be estimated. Estimating unobserved neighborhood structure is challenging, but Schweinberger (2017) showed that it can be recovered with high probability as long as the random graph is not too sparse and scaling and smoothness conditions are satisfied.

The aforementioned theoretical results suggest that statistical inference for exponential-family random graph models with additional structure is sensible, in contrast to statistical inference for exponential-family random graph models without additional structure. However, while these theoretical results are encouraging, it turns out that exponential-family random graph models with additional structure cannot be estimated from large networks, because the Bayesian methods of Schweinberger and Handcock (2015) cannot be applied to networks with more than one hundred nodes.

This thesis addresses these computational issues by proposing the first massive-scale estimation methods. The key idea is that the additional structure in the form of neighborhood structure is exploited to decompose random graphs into independent subgraphs. This gives rise to the following two-step likelihood-based approach: the first step estimates the neighborhood structure underlying the random graph by using variational approximations of the likelihood function supported by theoretical results. The second step estimates parameters given the estimated structure by computing the separate contributions of neighborhoods to the loglikelihood function. Both steps can be implemented in parallel and take advantage of computing clusters, which facilitates large-scale estimation of random graphs. The performance of the proposed model is assessed by a simulation study with both small and large networks. An application to a large real-world network with more than ten thousand nodes is presented to demonstrate the advantages of the two-step likelihood-based approach.

Chapter 2

High-dimensional multivariate time series with additional structure

Abstract

Chapter 2 discusses modeling of high-dimensional and dependent multivariate time series using vector autoregressive processes. While statistical inference for high-dimensional and dependent data is challenging, often additional structure in the form of spatial structure is available for many real-world data sets, e.g., geographical locations of monitors measuring air pollution. If it is known that a particular type of air pollution is short-range, the estimation of dependencies in a multivariate time series of pollution measurements can be restricted to pairs of components that are close to each other geographically. Hence, a novel two-step approach is proposed which estimates the range of dependence along with the parameters of vector autoregressive processes in high-dimensional settings. Additionally, this chapter provides non-asymptotic bounds on the statistical error of parameter estimates in high-dimensional settings and shows that the proposed approach reduces the statistical error. An application to air pollution in the U.S.A. demonstrates that the estimation approach reduces both computing time and prediction error. Moreover, it gives rise to results that are meaningful from a scientific point of view, in contrast to high-dimensional methods that ignore spatial structure and are less interpretable.

The contents of Chapter 2 have been accepted by the Journal of Computational and Graphical Statistics, which is published by the American Statistical Association: Schweinberger, M., Babkin, S., and K. B. Ensor (2017). High-dimensional multivariate time series with additional structure. Journal of Computational and Graphical Statistics.

2.1 Introduction

Multivariate time series (e.g., Lütkepohl, 2005; Wilson et al., 2015) arise in a wide range of applications, from finance to studies of air pollution and ecological studies (e.g., Ensor et al., 2013; Hoek et al., 2013; Chen et al., 2015). The age of computing has made it possible to collect data sets with large numbers of time series, where the number of parameters may exceed the number of observations. A common approach to dealing with high-dimensional data is to endow models with additional structure in the form of sparsity (e.g., Bühlmann and van de Geer, 2011). In the case of high-dimensional multivariate time series, an additional challenge is the complex dependence within and between time series. Some consistency results on model estimation and selection of high-dimensional vector autoregressive processes were obtained by Song and Bickel (2011), though under strong assumptions. Loh and Wainwright (2012) and Basu and Michailidis (2015) developed powerful concentration inequalities that enabled them to establish consistency under weaker conditions and prove that these conditions hold with high probability. In particular, Basu and Michailidis (2015) established consistency of ℓ_1-penalized least squares and maximum likelihood estimators of the autoregressive coefficients of high-dimensional vector autoregressive processes and related the estimation and prediction error to the complex dependence structure of vector autoregressive processes. Other estimation approaches, including Bayesian approaches, are discussed by Nguyen et al. (2014) and Davis et al. (2016).

We consider high-dimensional vector autoregressive processes with $p \gg N$ parameters, where p is the number of parameters and N is the number of observations. While high-dimensional vector autoregressive processes are challenging due to the dependent and high-dimensional nature of the data, in many applications there is additional structure that can be exploited to reduce computing time along with statistical error. Examples are studies of air pollution and ecological studies, where spatial structure can help reduce computing time and statistical error. If such structure is ignored, high-dimensional methods can give rise to results that contradict science. An example are daily measurements of ozone recorded by monitors across the U.S.A. as described in Section 2.6. Figure 2.1 shows the non-zero pattern of autoregressive coefficients estimated by the ℓ_1-penalized least

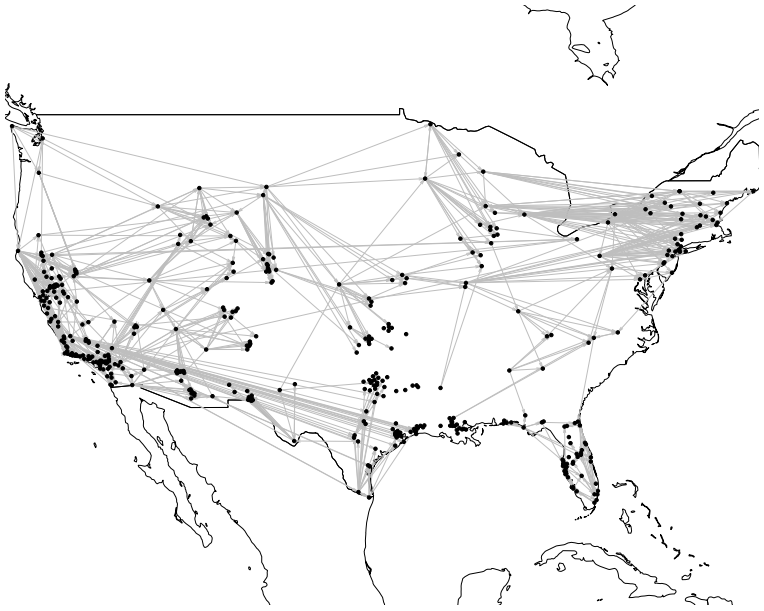


Figure 2.1 : Air pollution in the U.S.A.: autoregressive coefficients estimated by the ℓ_1-penalized least squares method from daily measurements of ozone. Monitors are connected by edges if the estimates of the corresponding autoregressive coefficients are non-zero. The long-distance edges contradict scientific evidence (see, e.g., Rao et al., 1997).

squares method described in Section 2.3.1. The figure suggests that today's ozone levels on the East Coast can directly affect tomorrow's ozone levels on the West Coast. Such results contradict science, because ozone cannot travel long distances (see, e.g., Rao et al., 1997).

We introduce novel methods and theory that take advantage of additional structure in the form of space with a view to reducing computing time along with statistical error, without making model assumptions about how the distance between the components of the vector autoregressive process affects the dependence between the components. We provide non-asymptotic bounds on the statistical error of parameter estimators in high-dimensional settings and show that the proposed approach reduces the statistical error. An application to air pollution recorded by 444 monitors across the U.S.A. with $N = 1{,}826$ observations and $p = 197{,}136$ parameters demonstrates that the proposed methods reduce both computing time and prediction error compared with existing high-dimensional methods and give rise to results that are meaningful from a scientific point of view, in contrast to high-dimensional methods that ignore the spatial structure. In practice, these high-

dimensional methods can be used to decompose high-dimensional multivariate time series into lower-dimensional multivariate time series that can be studied by other methods in more depth.

This chapter is structured as follows. We introduce vector autoregressive processes in Section 2.2. Methods and theory are described in Sections 2.3 and 2.4, respectively, followed by simulation results in Section 2.5 and an application in Section 2.6.

2.2 High-dimensional vector autoregressive processes with additional structure

We assume that $\boldsymbol{X}(t) = (X_1(t), \ldots, X_k(t))_{t=1}^N$ is generated by a L-th order vector autoregressive process of the form

$$\boldsymbol{X}(t) \;\; = \;\; \sum_{l=1}^{L} \boldsymbol{A}_l \, \boldsymbol{X}(t-l) + \boldsymbol{e}(t),$$

where $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L$ are $k \times k$ transition matrices and the errors $\boldsymbol{e}(t)$ are independent multivariate Gaussian random variables with mean $\mathbf{0}_k$ and positive-definite variance-covariance matrix $\boldsymbol{\Sigma}$. We follow Loh and Wainwright (2012) and Basu and Michailidis (2015) and assume that the order L of the vector autoregressive process is either known or can be bounded above and that the vector autoregressive process is stable and thus stationary (Lütkepohl, 2005). In applications where the order L of the vector autoregressive process is unknown and cannot be bounded above, cross-validation can be used to select L.

2.2.1 Additional structure

We consider high-dimensional vector autoregressive processes where the number of parameters $p = k^2 \, L + k^2$ is much larger than the number of observations N. While high-dimensional vector autoregressive processes are challenging due to the dependent and high-dimensional nature of the data, in many applications there is additional structure that can be exploited to reduce computing time along with statistical error. We consider high-dimensional vector autoregressive processes with additional structure in the form of space. In particular, we assume that the components i of

the vector autoregressive process have positions in the interior of a bounded subset $\mathbb{Z} \subset \mathbb{R}^d$. The boundedness assumption is motivated by applications: most spatial structures arising in applications can be represented by bounded subsets of $\mathbb{R}^d$. Throughout, we represent the components of the vector autoregressive process by a mixed graph, where the nodes represent components, a directed edge from component i to component j indicates that element (j, i) of at least one of the transition matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L$ is non-zero, and an undirected edge between components i and j indicates that elements (i, j) and (j, i) of $\boldsymbol{\Sigma}^{-1}$ are non-zero (Eichler, 2012). We note that the graphical representation of the model is convenient, but not essential: all results reported here could be described in terms of non-zero parameters.

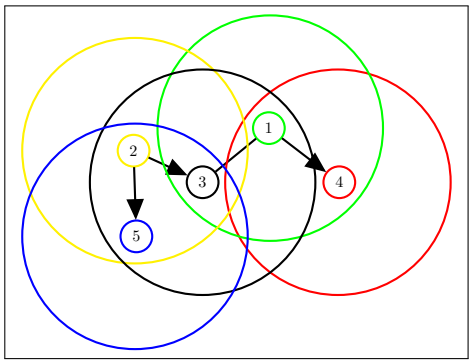
2.2.2 Model estimation exploiting additional structure

If additional structure is available, such as spatial structure, model estimation should take advantage of it.

To do so, observe that the boundedness of $\mathbb{Z} \subset \mathbb{R}^d$ implies that there exists $\rho_{\max} < \infty$ such that the Euclidean distance $d(i, j)$ between components i and j satisfies $d(i, j) \leq \rho_{\max}$ for all $(i, j) \in \mathcal{N} \times \mathcal{N}$, where $\mathcal{N} = \{1, \ldots, k\}$ denotes the set of components. Let ρ be the maximum distance separating two components (i, j) with an edge. By definition of ρ, for each component i, all edges of i are either in the interior or on the boundary of the closed ball centered at the position of i in $\mathbb{Z} \subset \mathbb{R}^d$ with radius $\rho \leq \rho_{\max}$ (see, e.g., Figure 2.2). In light of the fact that all edges, i.e., all non-zero parameters of all components are within distances $d \leq \rho$, model estimation of non-zero parameters should be restricted to distances $d \leq \rho$.

In practice, the radius ρ is sometimes known or can be bounded above based on domain knowledge, but in most cases ρ is unknown and must be estimated. We introduce methods and theory for estimating ρ in Sections 2.3 and 2.4 with a view to reducing computing time along with the statistical error of parameter estimators. It is worth noting that we do not make model assumptions about how the distance between components of the vector autoregressive process affects the dependence between the components: all we assume is that components have positions in a bounded

$$A_1 = \begin{pmatrix} \star & 0 & 0 & 0 & 0 \\ 0 & \star & 0 & 0 & 0 \\ 0 & \star & \star & 0 & 0 \\ \star & 0 & 0 & \star & 0 \\ 0 & \star & 0 & 0 & \star \end{pmatrix}$$



$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \star & 0 & \star & 0 & 0 \\ 0 & \star & 0 & 0 & 0 \\ \star & 0 & \star & 0 & 0 \\ 0 & 0 & 0 & \star & 0 \\ 0 & 0 & 0 & 0 & \star \end{pmatrix}$$

Figure 2.2 : First-order vector autoregressive process with additional structure: nodes represent components of the vector autoregressive process with positions in a bounded subset $\mathbb{Z} \subset \mathbb{R}^d$ and edges represent non-zero elements of either $\boldsymbol{A}_1$ or $\boldsymbol{\Sigma}^{-1}$. The edges of components are contained in the closed balls with radius ρ centered at the positions of the components. The elements $\star$ of matrices indicate non-zero elements.

subset $\mathbb{Z} \subset \mathbb{R}^d$. Therefore, the methods can be applied to all vector autoregressive processes with additional structure of the form considered here, including vector autoregressive processes with $\rho = \rho_{\max}$, but the greatest reduction in computing time and statistical error is obtained when $\rho \ll \rho_{\max}$ and the components are not too close to each other in $\mathbb{Z} \subset \mathbb{R}^d$.

2.3 Two-step ℓ_1-penalized least squares method

We introduce a simple two-step ℓ_1-penalized least squares method that takes advantage of the additional structure considered here.

The two-step ℓ_1-penalized least squares method is sketched in Table 2.1. It is motivated by the fact that all edges, i.e., all non-zero parameters of all components are within distances $d \leq \rho$, thus model estimation of non-zero parameters should be restricted to distances $d \leq \rho$. In practice, the radius ρ may be unknown. If the structure of the graph was known, one could take ρ to be the maximum distance that separates a pair of nodes with an edge. If the structure of the graph is unknown, one needs to estimate the graph. An appealing alternative to estimating the whole graph—which is time-consuming when the set of nodes $\mathcal{N}$ is large—is to estimate a subgraph by

1. If radius ρ is unknown, estimate ρ:
   - 1.1 Sample a subset of nodes $\mathbb{S}$ from the set of nodes $\mathcal{N}$.
   - 1.2 Estimate edges by regressing nodes $i \in \mathbb{S}$ on $\{j \mid j \in \mathcal{N} \setminus i\}$, i.e., on all other nodes in $\mathcal{N}$.
   - 1.3 Estimate radius ρ by $\widehat{\rho}$, the maximum distance that separates a pair of nodes with an estimated edge.
2. Estimate the parameters by using the ℓ_1-penalized least squares method subject to the constraint that all parameters governing possible edges at distances $d > \widehat{\rho}$ are 0.

Table 2.1 : Two-step ℓ_1-penalized least squares method.

sampling a subset of nodes $\mathbb{S}$, estimating the edges of nodes $i \in \mathbb{S}$, and then estimating ρ by $\widehat{\rho}$, defined as the maximum distance that separates a pair of nodes with an estimated edge. Step 1 estimates the radius ρ by $\widehat{\rho}$ along these lines. Step 2 estimates the parameters by restricting the estimation of parameters to distances $d \leq \widehat{\rho}$. If the sample in Step 1 is small but well-chosen and the radius ρ is small, the two-step ℓ_1-penalized least squares method reduces computing time and statistical error.

We discuss the implementation of the two-step ℓ_1-penalized least squares method in Sections 2.3.1 and 2.3.2 and shed light on its theoretical properties in Section 2.4. Throughout, we assume that $\mathbf{\Sigma}^{-1}$ is diagonal; extensions to non-diagonal $\mathbf{\Sigma}^{-1}$ are possible, though less attractive on computational grounds (Basu and Michailidis, 2015). We denote by $\|.\|_1$, $\|.\|_2$, and $\|.\|_\infty$ the ℓ_1, ℓ_2, and ℓ_∞-norm of vectors, respectively. The total number of observations is denoted by M and the effective number of observations by $N = M - L + 1$.

2.3.1 Step 1

If the radius ρ is unknown, it is estimated in Step 1.

In Step 1.1, a sample of nodes $\mathbb{S}$ from the set of nodes $\mathcal{N}$ is generated by using any sampling design for sampling from finite populations (see, e.g., Thompson, 2012). Some guidance with respect to sampling designs is provided in Remark 7 in Section 2.4. An example is given in Section 2.6.

In Step 1.2, edges are estimated by regressing nodes $i \in \mathbb{S}$ on $\{j \mid j \in \mathcal{N} \setminus i\}$ by the ℓ_1-penalized least squares method of Basu and Michailidis (2015), which is attractive on both computational and theoretical grounds. It is worth noting that regressing sampled nodes $i \in \mathbb{S}$ on all other sampled nodes in $\mathbb{S}$ would give rise to omitted variable problems. Step 1.2 therefore regresses sampled nodes $i \in \mathbb{S}$ on *all other nodes in* $\mathcal{N}$ rather than *all other sampled nodes in* $\mathbb{S}$.

To introduce the ℓ_1-penalized least squares method used in Step 1.2, note that the conventional ℓ_1-penalized least squares method estimates the $p = k^2\,L$-dimensional parameter vector $\boldsymbol{\beta}_{\mathcal{N}} = (\boldsymbol{\beta}_i)_{i\in\mathcal{N}}$ corresponding to the vectorized transition matrices $\text{vec}(\boldsymbol{A}_1^\top, \ldots, \boldsymbol{A}_L^\top)$ by

$$\widehat{\boldsymbol{\beta}}_{\mathcal{N}} \;\in\; \underset{\boldsymbol{\beta}_i,\, i\in\mathcal{N}}{\arg\min} \sum_{i\in\mathcal{N}} \left[\frac{1}{N}\|\boldsymbol{\mathcal{Y}}_i - \boldsymbol{\mathcal{X}}\,\boldsymbol{\beta}_i\|_2^2 + \lambda_1\,\|\boldsymbol{\beta}_i\|_1\right], \tag{2.1}$$

where $\boldsymbol{\beta}_i$ denotes the $p_i = k\,L$-dimensional parameter vectors governing possible incoming edges of nodes i; $\boldsymbol{\mathcal{Y}}_i$ denotes the i-th column of the matrix of observations $\boldsymbol{\mathcal{Y}} = (\boldsymbol{X}(M)^\top, \ldots, \boldsymbol{X}(L)^\top)$; $\boldsymbol{\mathcal{X}}$ denotes the predictors $((\boldsymbol{X}(M-1)^\top, \ldots, \boldsymbol{X}(L-1)^\top), \ldots, (\boldsymbol{X}(M-L)^\top, \ldots, \boldsymbol{X}(0)^\top))$; and $\lambda_1 > 0$ denotes a regularization parameter. The ℓ_1-penalized least squares method used in Step 1.2 applies the same procedure to the subset of nodes $\mathbb{S}$ and estimates the parameter vector $\boldsymbol{\beta}_{\mathbb{S}} = (\boldsymbol{\beta}_i)_{i\in\mathbb{S}}$ by

$$\widehat{\boldsymbol{\beta}}_{\mathbb{S}} \;\in\; \underset{\boldsymbol{\beta}_i,\, i\in\mathbb{S}}{\arg\min} \sum_{i\in\mathbb{S}} \left[\frac{1}{N}\|\boldsymbol{\mathcal{Y}}_i - \boldsymbol{\mathcal{X}}\,\boldsymbol{\beta}_i\|_2^2 + \lambda_1\,\|\boldsymbol{\beta}_i\|_1\right]. \tag{2.2}$$

The incoming edges of nodes $i \in \mathbb{S}$ can be inferred from the non-zero pattern of $\widehat{\boldsymbol{\beta}}_{\mathbb{S}} = (\widehat{\boldsymbol{\beta}}_i)_{i\in\mathbb{S}}$. The radius ρ can be estimated by $\widehat{\rho}$, the maximum distance that separates a pair of nodes $(j,i) \in \mathcal{N}\times\mathbb{S}$ with an estimated edge, i.e., with an estimated non-zero autoregressive coefficient.

2.3.2 Step 2

In Step 2, the parameter vector $\boldsymbol{\beta} \equiv \boldsymbol{\beta}_{\mathcal{N}}$ is estimated by restricting the ℓ_1-penalized least squares method to distances $d \leq \widehat{\rho}$, i.e., the parameter vector $\boldsymbol{\beta}$ is estimated by

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}_i, i \in \mathcal{N}}{\arg\min} \sum_{i \in \mathcal{N}} \left[\frac{1}{N} \| \boldsymbol{\mathcal{Y}}_i - \boldsymbol{\mathcal{X}} \, \boldsymbol{\beta}_i \|_2^2 + \lambda_2 \, \| \boldsymbol{\beta}_i \|_1 \right] \tag{2.3}$$

subject to the constraint that all parameters governing possible edges at distances $d > \widehat{\rho}$ are 0, where $\lambda_2 > 0$ is a regularization parameter.

Remark 1. An important observation is that the parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ are variation-independent in the sense that the parameter space of $\boldsymbol{\beta}$ is a product space of the form $\mathbb{R}^{k^2 L} = \mathbb{R}^{kL} \times \cdots \times \mathbb{R}^{k\,L}$. As a result, optimization problems (2.1), (2.2), and (2.3) can be decomposed into k separate optimization problems that can be solved in parallel, thus reducing computing time.

Remark 2. The variance-covariance matrix $\boldsymbol{\Sigma}$ can be estimated by using the ℓ_1-penalized maximum likelihood method of Basu and Michailidis (2015). However, the ℓ_1-penalized maximum likelihood method is more expensive in terms of computing time than the ℓ_1-penalized least squares method.

2.4 Theoretical properties

We provide non-asymptotic bounds on the statistical error of parameter estimators in high-dimensional settings and show that the two-step ℓ_1-penalized least squares method reduces the statistical error. To facilitate the discussion, we follow Loh and Wainwright (2012) and Basu and Michailidis (2015) by expressing optimization problems (2.1), (2.2), and (2.3) as M-estimation problems of the form

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{C}}{\arg\min} \left[-2 \, \boldsymbol{\beta}^\top \widehat{\boldsymbol{\gamma}} + \boldsymbol{\beta}^\top \, \widehat{\boldsymbol{\Gamma}} \, \boldsymbol{\beta} + \lambda \, \| \boldsymbol{\beta} \|_1 \right],$$

where $\mathbb{C}$ is a subset of $\mathbb{R}^p$ that depends on the constraints imposed by optimization problems (2.1), (2.2), and (2.3), $\widehat{\boldsymbol{\gamma}} = (\boldsymbol{I} \otimes \boldsymbol{\mathcal{X}}^\top) \, \mathrm{vec}(\boldsymbol{\mathcal{Y}})/N$, and $\widehat{\boldsymbol{\Gamma}} = (\boldsymbol{I} \otimes \boldsymbol{\mathcal{X}}^\top \boldsymbol{\mathcal{X}})/N$, where $\boldsymbol{I}$ denotes the identity matrix of suitable order and $\otimes$ denotes the Kronecker product.

Notation. Throughout, we assume that the elements of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are ordered according to distance and denote by $\boldsymbol{\beta}_{[d_1,d_2]}$ and $\boldsymbol{\gamma}_{[d_1,d_2]}$ the subvectors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ corresponding to parameters governing possible edges at distances $d \in [d_1, d_2]$, respectively, where $0 \leq d_1 \leq d_2$. The rows and columns of $\boldsymbol{\Gamma}$ are ordered in accordance. Denote by $p(0, d_2)$ the total number of parameters governing possible edges at distances $d \in [0, d_2]$ and by $p(d_1, d_2)$ the total number of parameters governing possible edges at distances $d \in (d_1, d_2]$, where $0 < d_1 \leq d_2$. Let $\widehat{\boldsymbol{\beta}}$ be the estimator of the true parameter vector $\boldsymbol{\beta}^\star$ obtained by the two-step ℓ_1-penalized least squares method. Denote by $\mathbb{S}$ the support of $\boldsymbol{\beta}^\star$ and by s the size of support $\mathbb{S}$. Let $\delta > 0$ and $\mathbb{S}(\delta)$ be the subset of nodes with incoming edges at distances $d \in [\rho - \delta, \rho]$. We denote by $c_0, c_1, c_2 > 0$ unspecified constants.

We assume that the following conditions hold. The first assumption is a restricted eigenvalue condition, whereas the second condition is a deviation condition. Both conditions are conventional and hold with high probability (Loh and Wainwright, 2012; Basu and Michailidis, 2015).

Condition C.1. $\widehat{\boldsymbol{\Gamma}}$ satisfies the restricted eigenvalue condition with curvature $\alpha > 0$ and tolerance $\tau > 0$ provided $s\,\tau \leq \alpha/32$ and

$$\boldsymbol{b}^\top \widehat{\boldsymbol{\Gamma}}\, \boldsymbol{b} \;\; \geq \;\; \alpha \, \|\boldsymbol{b}\|_2^2 - \tau \, \|\boldsymbol{b}\|_1^2 \quad \text{for all} \quad \boldsymbol{b} \in \mathbb{R}^p.$$

Condition C.2. There exists a deterministic function $\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) > 0$ such that $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$ satisfy

$$\|\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\, \boldsymbol{\beta}^\star\|_\infty \;\; \leq \;\; \mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) \sqrt{\frac{\log p}{N}}.$$

The following theorems show that the two-step ℓ_1-penalized least squares method reduces the statistical error of parameter estimators without making model assumptions about how the distance between the components of the vector autoregressive process affects the dependence between the components. We start with the case where ρ is either known or can be bounded above based on domain knowledge (Theorem 2.1) and then turn to the case of unknown ρ (Theorem 2.2). To streamline the presentation, Theorem 2.1 focuses on known ρ, but the extension to bounded ρ is straightforward.

Theorem 2.1 *Consider $N \geq c_0\, s \log p$ ($c_0 > 1$) observations from a stable L-th order vector autoregressive process with radius $\rho > 0$. Suppose that ρ is known and that the regularization parameter λ_2 in the second step of the two-step ℓ_1-penalized least squares method satisfies*

$$\lambda_2 \;\; \geq \;\; 4\, \mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) \sqrt{\frac{\log p(0,\rho)}{N}}. \tag{2.4}$$

Then, with at least probability

$$1 - 2\, \exp(-c_1\, N) - 6\, \exp(-c_2 \log p(0,\rho)), \tag{2.5}$$

the ℓ_2-error of estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^\star$ is bounded above by

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \;\; \leq \;\; \frac{16\, \sqrt{s}\, \lambda_2}{\alpha}.$$

We compare the statistical error and computing time of the two-step ℓ_1-penalized least squares method to existing high-dimensional methods.

Remark 3. Comparison in terms of statistical error. Among the existing approaches, the most attractive approach is the ℓ_1-penalized least squares method of Basu and Michailidis (2015), because it has computational advantages and its theoretical properties are well-understood. Suppose that $\boldsymbol{\beta}^\star$ is estimated by the two-step ℓ_1-penalized least squares method with known ρ with $\lambda_2 = 4\, \mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) \sqrt{\log p(0,\rho)/N}$. Then, with high probability,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \;\; \leq \;\; \underbrace{\frac{64}{\alpha} \mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) \sqrt{\frac{s \log p(0,\rho)}{N}}}_{\text{two-step } \ell_1\text{-least squares}} \;\; \leq \;\; \underbrace{\frac{64}{\alpha} \mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma}) \sqrt{\frac{s \log p}{N}}}_{\ell_1\text{-least squares}},$$

because $p(0,\rho) = \sum_{i=1}^{k} n_i(\rho)\, L \leq p = k^2\, L$, where $p(0,\rho)$ is the total number of parameters governing possible edges at distances $d \in [0,\rho]$ and $n_i(\rho)$ is the number of components $j \in \mathcal{N} \setminus i$ within distance $d(i,j) \leq \rho$ of component i. The error bounds show that restricting model estimation to distances $d \leq \rho$ reduces the ℓ_2-error of $\widehat{\boldsymbol{\beta}}$.

Remark 4. Comparison in terms of computing time. In terms of computing time, the two-step ℓ_1-penalized least squares method with known (bounded) ρ tends to be superior to the ℓ_1-penalized least squares method: while the ℓ_1-penalized least squares method amounts to running

k regressions with $k\,L$ predictors, the two-step ℓ_1-penalized least squares method with known (bounded) ρ amounts to running k regressions with $\max_{1\leq i\leq k} n_i(\rho)\,L$ predictors, where $n_i(\rho)$ is the number of components $j \in \mathcal{N}\setminus i$ within distance $d(i,j) \leq \rho$ of component i. If $\max_{1\leq i\leq k} n_i(\rho) \ll k$, the two-step ℓ_1-penalized least squares method with known (bounded) ρ is much faster than the ℓ_1-penalized least squares method and can thus be applied to much larger data sets.

We turn to the case where ρ is unknown. Choose $\delta > 0$ as small as desired and consider the estimator $\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}$ of the parameter vector $\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}$ governing possible edges to nodes in the interior of the balls centered at the positions of nodes, which—in most applications—are the parameters of primary interest. Theorem 2.2 bounds the statistical error of the estimator $\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}$ of the parameter vector $\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}$ for all $\delta > 0$.

Theorem 2.2 *Consider $N \geq c_0\, s \log p$ ($c_0 > 1$) observations from a stable L-th order vector autoregressive process with radius $\rho > 0$. Assume that components i are sampled independently with probabilities $0 < \theta_i < 1$ and that the minimum signal strength is $\beta^{\star}_{\min} = \min_{i\in\mathbb{S}} |\beta^{\star}_i| \geq 32\,\sqrt{s}\,\lambda_1/\alpha > 0$. Choose any $\delta > 0$, however small, and assume that the regularization parameters λ_1 and λ_2 in the first and second step of the two-step ℓ_1-penalized least squares method satisfy*

$$\lambda_1 \;\geq\; 4\,\mathbb{Q}(\boldsymbol{\beta}^{\star},\boldsymbol{\Sigma})\,\sqrt{\frac{\log p}{N}} \tag{2.6}$$

and

$$\lambda_2 \;\geq\; 4\,\mathbb{Q}(\boldsymbol{\beta}^{\star},\boldsymbol{\Sigma})\,\sqrt{\frac{\log p(0,\rho-\delta)}{N}}, \tag{2.7}$$

respectively. Then, for all $\delta > 0$, with at least probability

$$1 - 4\exp(-c_1 N) - 12\exp(-c_2 \log p(0,\rho-\delta)) - \exp\left(-\sum_{i\in\mathbb{S}(\delta)} \theta_i\right), \tag{2.8}$$

the ℓ_2-error of the estimator $\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}$ of the parameter vector $\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}$ is bounded above by

$$\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_2 \;\leq\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha}.$$

Remark 5. Statistical error. The so-called beta-min condition in Theorem 2.2, which asserts that the non-zero elements of $\boldsymbol{\beta}^\star$ cannot be too small, is common in the literature on high-dimensional variable selection and graphical models (see, e.g., Bühlmann and van de Geer, 2011, Section 7.4). It is needed to make sure that the edges of sampled nodes can be recovered with high probability, which in turn is needed to estimate the radius ρ. Theorem 2.2 shows that the statistical error of estimators of the parameter vector $\boldsymbol{\beta}^\star_{[0,\rho-\delta]}$ governing possible edges in the interior of the balls—which, in most applications, are the parameters of primary interest—is small when the number of observations N is large relative to the size of the support s and the number of parameters $p(0, \rho - \delta)$. The statistical error of the estimator $\widehat{\boldsymbol{\beta}}$ of the whole parameter vector $\boldsymbol{\beta}^\star$ is more complicated. On the one hand, if ρ is overestimated in Step 1, the error bound of the estimator $\widehat{\boldsymbol{\beta}}$ in Step 2 is at most as large as the error bound of the estimator $\widehat{\boldsymbol{\beta}}$ under the ℓ_1-penalized least squares method, which follows from Theorem 2.1 and Remark 3. On the other hand, if ρ is underestimated in Step 1, the parameter vector $\boldsymbol{\beta}^\star_{(\rho-\delta,\rho]}$ governing possible edges close to the boundary of the balls centered at the positions of nodes is not estimated in Step 2, thus the error bound of the estimator $\widehat{\boldsymbol{\beta}}$ in Step 2 depends on the ℓ_2-norm of $\boldsymbol{\beta}^\star_{(\rho-\delta,\rho]}$.

Remark 6. Computing time. In terms of computing time, the two-step ℓ_1-penalized least squares method amounts to running $|\mathbb{S}|$ regressions with $k\,L$ predictors in Step 1 and k regressions with $\max_{1\leq i\leq k} n_i(\widehat{\rho}\,)\,L$ predictors in Step 2 of the two-step ℓ_1-penalized least squares method, where $\widehat{\rho}$ is the estimate of ρ obtained in Step 1. Therefore, as long as the sample is small but well-chosen and the radius is short, the two-step ℓ_1-penalized least squares method outperforms the ℓ_1-penalized least squares method.

Remark 7. Sampling. Theorem 2.2 shows that, for any given $\delta > 0$, the probability of the event that $\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_2$ is small depends on the term $\exp(-\sum_{i\in\mathbb{S}(\delta)}\theta_i)$: that is, it depends on (a) the size of $\mathbb{S}(\delta)$ and (b) the sample inclusion probabilities θ_i of nodes $i \in \mathbb{S}(\delta)$, i.e., nodes with incoming edges at distances $d \in [\rho - \delta, \rho]$. The first factor is outside of the control of investigators, whereas the second factor is under the control of investigators. The fact that the probability of the event of interest depends on the sample inclusion probabilities θ_i of nodes $i \in$

$\mathbb{S}(\delta)$ rather than nodes $i \in \mathcal{N} \setminus \mathbb{S}(\delta)$ shows that one needs to sample nodes $i \in \mathbb{S}(\delta)$ rather than nodes $i \in \mathcal{N} \setminus \mathbb{S}(\delta)$ with high probability. In other words, non-uniform sampling designs that sample nodes with long-distance edges with high probability are preferable to uniform sampling designs and the number of sampled nodes with long-distance edges is more important than the total number of sampled nodes. Therefore, if prior knowledge is available about which nodes may have long-distance edges, it should be incorporated into the sampling design. Such prior knowledge is available in a number of spatio-temporal applications: e.g., in studies of air pollution, it is well-known that industrial and metropolitan areas tend to spread air pollution to surrounding areas and that some geographical conditions in combination with wind conditions facilitate long-distance transport of pollutants. Thus, pollution monitors in industrial and metropolitan areas and other areas suspected of facilitating long-distance transport of pollutants should be sampled with high probability. In the application in Section 2.6, we sample pollution monitors in the 15 most polluted cities in the U.S. with high probability and others with low probability.

2.5 Simulation results

We compare the two-step ℓ_1-penalized least squares method with known ρ and unknown ρ to the ℓ_1-penalized least squares method of Basu and Michailidis (2015), which is the most attractive high-dimensional method available, as discussed in Remark 3 in Section 2.4. We compare the methods in terms of statistical error and computing time. Throughout, we use stability selection (Meinshausen and Bühlmann, 2010) to sidestep the problem that the choice of the regularization parameters λ_1 and λ_2 in the first and second step of the two-step ℓ_1-penalized estimation method depends on the unknown values of $\boldsymbol{\beta}^\star$ and $\boldsymbol{\Sigma}$. We followed the guidelines of Meinshausen and Bühlmann (2010) concerning the choice of tuning parameters of stability selection. The `R` source code we used is contained in the supplementary archive.

To shed light on the statistical error of the methods, we consider three high-dimensional scenarios with $N = 150$ ($k = 100$), $N = 300$ ($k = 200$), and $N = 450$ ($k = 300$) observations; note that $p = k^2\, L \gg N$ in all three cases. For each scenario, we generated data from a first-order vector

		$k = 100$	$k = 200$	$k = 300$
AUROC	Least squares	.994 (.005)	.968 (.013)	.867 (.033)
	Two-step least squares	.987 (.016)	.988 (.011)	.960 (.021)
	Oracle two-step least squares	.999 (.001)	.996 (.003)	.969 (.019)
Estimation error	Least squares	.374 (.026)	.525 (.032)	.714 (.043)
	Two-step least squares	.343 (.028)	.492 (.037)	.666 (.052)
	Oracle two-step least squares	.324 (.019)	.479 (.032)	.655 (.052)
Fraction of FP	Least squares	.003 (.000)	.003 (.001)	.005 (.000)
	Two-step least squares	.001 (.000)	.002 (.000)	.004 (.001)
	Oracle two-step least squares	.001 (.000)	.002 (.000)	.004 (.001)
Fraction of FN	Least squares	.054 (.016)	.105 (.028)	.291 (.058)
	Two-step least squares	.034 (.022)	.052 (.036)	.156 (.068)
	Oracle two-step least squares	.018 (.013)	.033 (.018)	.133 (.062)

Table 2.2 : Comparison of the ℓ_1-penalized least squares method, the two-step ℓ_1-penalized least squares method with unknown ρ, and the oracle two-step ℓ_1-penalized least squares method with known ρ. Monte Carlo standard deviations are given in parentheses.

autoregressive process with $k \times k$ transition matrix $\boldsymbol{A} \equiv \boldsymbol{A}_1$ with 2% sparsity and overlapping neighborhoods. The overlapping neighborhoods are generated as follows: we sample 5 ($k = 100$), 10 ($k = 200$), and 15 ($k = 300$) points from the Uniform distribution on a two-dimensional square. The sampled points are considered to be centers of neighborhoods and, for each neighborhood, we sample 20 points from a bivariate Gaussian centered at the neighborhood center. Then edges are generated so that 90% of all edges are within neighborhoods and 10% are between neighborhoods, subject to the constraint that between-neighborhood edges are at distances less than the 30% quantile of the empirical distribution of distances. We compare the methods in terms of (a) model selection error: the area under the receiving operator characteristic curve (AUROC); the fraction of false-positive (FP) and false-negative (FN) edges; and (b) model estimation error: the relative

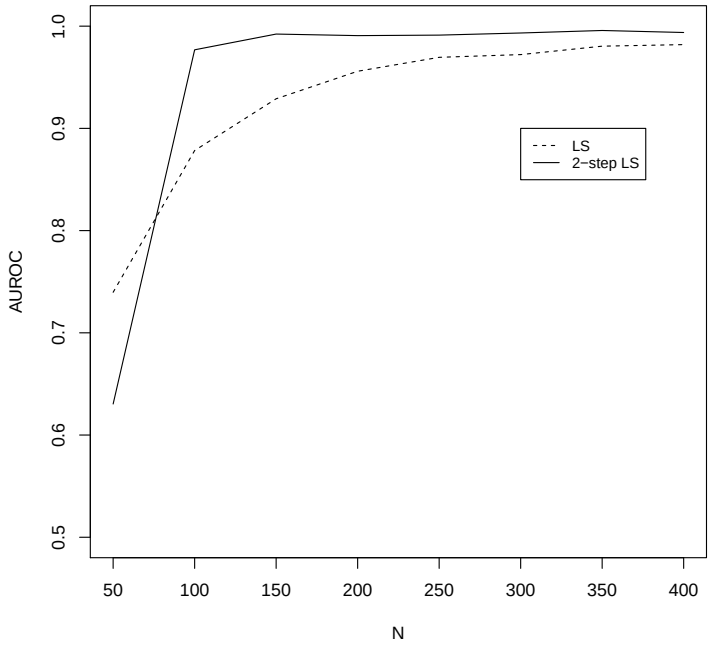


Figure 2.3 : AUROC plotted against number of observations N using $k = 200$ components. The dashed and solid line correspond to the ℓ_1-penalized least squares method (LS) and the two-step ℓ_1-penalized least squares method with unknown ρ (2-step LS), respectively.

estimation accuracy measured by $||\boldsymbol{A} - \widehat{\boldsymbol{A}}||_F/||\boldsymbol{A}||_F$, where $||\boldsymbol{A}||_F = \sqrt{\text{tr}(\boldsymbol{A}^\top \boldsymbol{A})}$. In Table 2.2, we report the results based on 1,000 Monte Carlo simulations along with Monte Carlo standard deviations. It is not surprising that the oracle two-step ℓ_1-penalized least squares method with known ρ seems to perform best, but the two-step ℓ_1-penalized least squares method with unknown ρ seems to be close. Both seem to outperform the ℓ_1-penalized least squares method. In Figure 2.3, we assess the impact of the number of observations N on model selection error in terms of AUROC using $k = 200$ components. It is evident that the two-step ℓ_1-penalized least squares method with unknown ρ outperforms the ℓ_1-penalized least squares method even when N is as small as 100.

To compare the methods in terms of computing time, we consider $k = 400$ time series governed by a first-order vector autoregressive process with a 400×400 transition matrix $\boldsymbol{A} \equiv \boldsymbol{A}_1$ with 1% sparsity in the high-dimensional setting where $p = 160{,}000 \gg N = 600$. To assess the

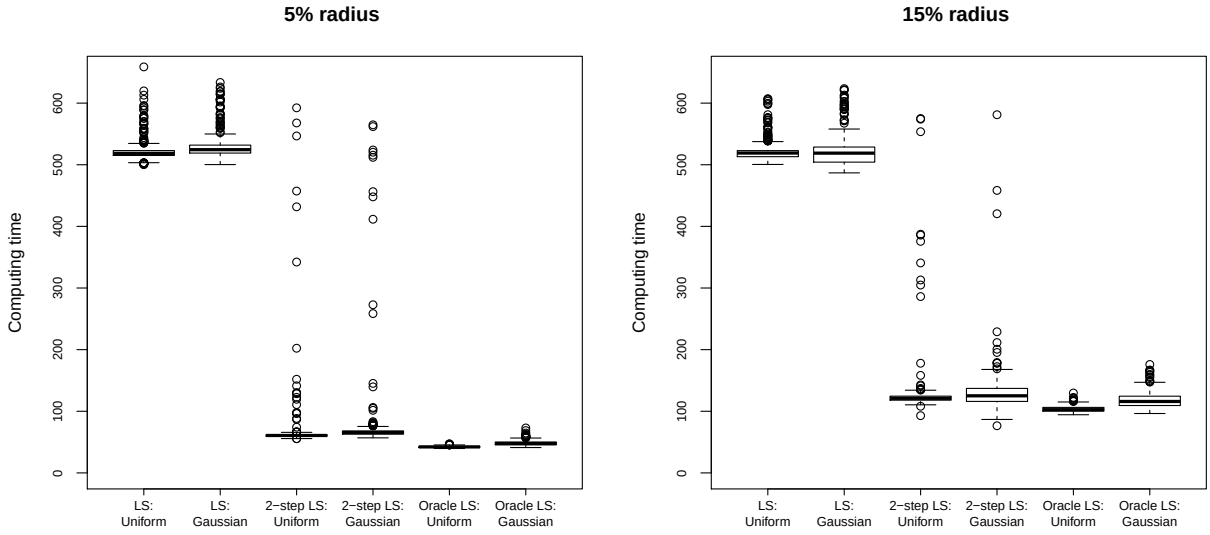


Figure 2.4 : Computing time in seconds of the ℓ_1-penalized least squares method (LS), the two-step ℓ_1-penalized least squares method with unknown ρ (2-step LS), and the oracle two-step ℓ_1-penalized least squares method with known ρ (Oracle LS) in two spatial settings (Uniform and Gaussian) with small and moderate radius (5% and 15%).

impact of the spatial structure and the radius on computing time, we compare the methods in two spatial settings and, for each spatial setting, we use a small and a moderate radius ρ. The two spatial settings are generated by two processes. The first generating process, called Uniform generating process, generates spatial positions of time series by sampling 400 points from the Uniform distribution on a two-dimensional square. The second generating process, called Gaussian generating process, generates spatial positions of time series by first sampling 20 points from the Uniform distribution on a two-dimensional square. The 20 points are used as centers of 20 bivariate Gaussians and from each bivariate Gaussian 20 points are sampled. For each spatial structure, we select a small and a moderate radius. To make sure that the balls centered at the locations of the time series contain a non-negligible fraction of possible edges, we use the 5% and 15% quantile of the empirical distribution of the distances as small and moderate radius, respectively. The corresponding radii are called “5% radius” and “15% radius”, but note that the resulting radius varies from data set to data set, depending on the spatial positions of the time series. Conditional

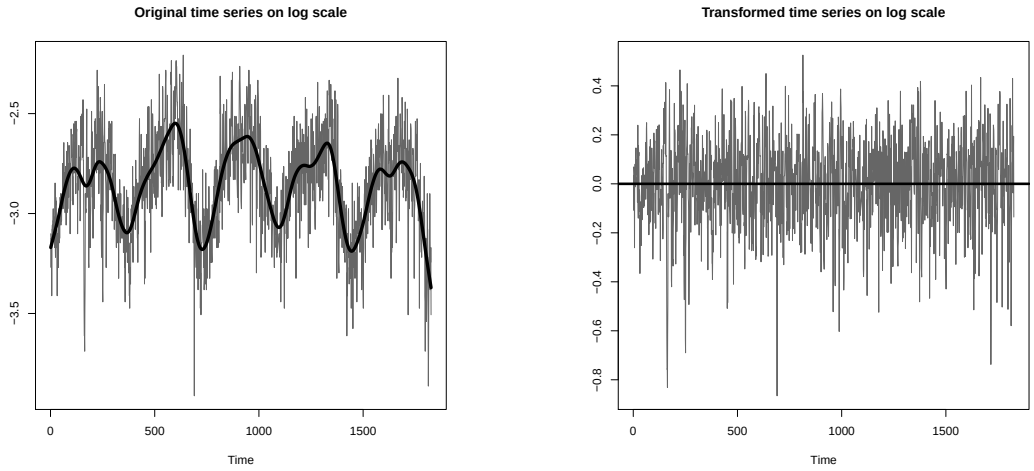


Figure 2.5 : Example of ozone time series consisting of $N = 1{,}826$ observations of ozone levels between January 2010 and December 2014 in its original form and transformed form, both on the log scale. The figure on the left-hand side shows the original log ozone time series. The 5 summers increase the log ozone levels while the 5 winters decrease them. The black curve is the fitted cubic spline that captures the seasonal ups and downs. The figure on the right-hand side shows the transformed log ozone time series. The black line is the mean of the $N = 1{,}826$ observations.

on the locations of the $k = 400$ time series, we generate edges by sampling pairs of time series without replacement to achieve 1% sparsity and then generate $N = 600$ observations from a first-order vector autoregressive process with 400×400 transition matrix $\boldsymbol{A} \equiv \boldsymbol{A}_1$ with 1% sparsity. The results based on 500 Monte Carlo simulations are presented in Figure 2.4. The figure demonstrates that the two-step ℓ_1-penalized least squares method with known ρ and unknown ρ outperforms the ℓ_1-penalized least squares method in terms of computing time by a factor of close to 10 (5% radius) and 5 (15% radius). The two-step ℓ_1-penalized least squares method with unknown ρ is almost as fast as the oracle version with known ρ, demonstrating that estimating ρ rather than knowing ρ comes at a cost, but the cost seems to be low, with the exception of the rare cases where ρ is overestimated by a non-negligible amount. The impact of the spatial structure on the computing time seems to be small, but increasing the radius increases the computing time visibly.

2.6 Application to air pollution in the U.S.A.

Air pollution is an important health concern. The American Lung Association (2015) states that in the U.S.A. alone almost 138.5 million people live in areas where air pollution makes breathing dangerous. Air pollution has been associated with cardiac arrest (Ensor et al., 2013), lung disease (Hoek et al., 2013), and cancer (Chen et al., 2015), and the World Health Organization (2014) attributed more than 7 million deaths in 2012 alone to air pollution.

We exploit the two-step ℓ_1-penalized least squares method to contribute to the understanding of the 24-hour transport of air pollution across space by using data from the U.S. Environmental Protection Agency obtained from

`http://www.epa.gov/airdata/ad_data_daily.html`.

The data are contained in the supplementary archive along with all `R` source code we used to analyze the data. Throughout the section, we use first-order vector autoregressive processes, because ozone and other pollutants tend to decompose fast. An additional advantage of using first-order vector autoregressive processes is that we have ground truth on the 24–72 hour transport of ozone in the sense that we have an upper bound on the spatial distance ozone is known to travel in 24–72 hours (see, e.g., Rao et al., 1997). We first take a bird's eye view at air pollution in the U.S.A. (Section 2.6.1) and then zoom in on the Gulf of Mexico region (Section 2.6.2), one of the most monitored regions in the U.S.A.

2.6.1 A bird's eye view: air pollution in the U.S.A.

We consider daily measurements of 8-hour maximum concentration of ozone (O_3) recorded by monitors across the U.S.A. The data set consists of $N = 1{,}826$ observations of ozone levels recorded by $k = 444$ monitors between January 2010 and December 2014. All monitors contain less than 10% of missing values and we impute the missing values by univariate linear interpolation. ozone concentrations were log-transformed and a cubic spline was fitted to each ozone time series to capture the seasonal ups and downs. We subtract the fitted cubic splines from the log-

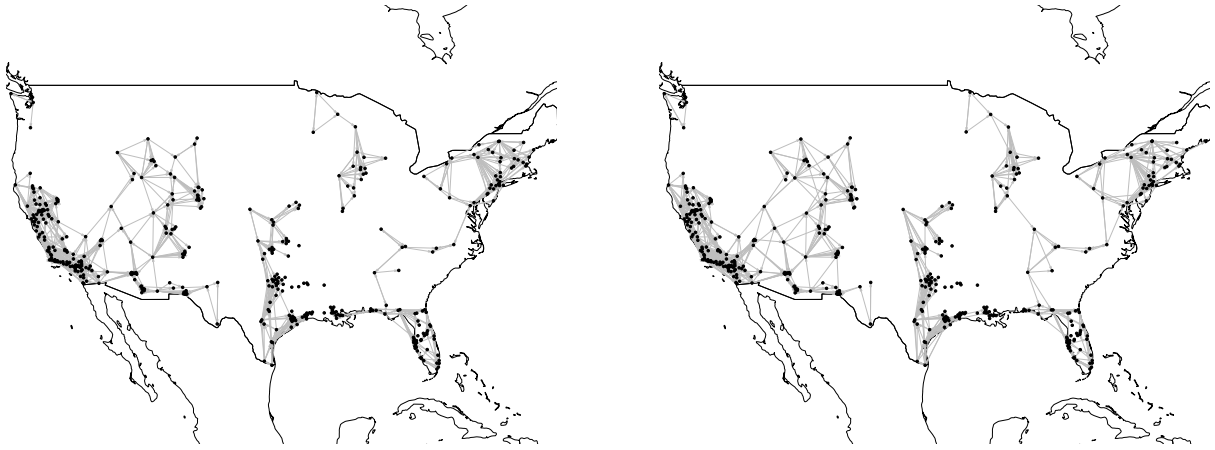


Figure 2.6 : Air pollution in the U.S.A.: autoregressive coefficients estimated by the two-step ℓ_1-penalized least squares method with estimate $\widehat{\rho} = 239$ (left) and upper bound $\rho = 250$ (right), where the upper bound is based on scientific evidence. Monitors are connected by edges if the estimates of the corresponding autoregressive coefficients are non-zero. The results demonstrate that the two-step ℓ_1-penalized least squares method respects the fact that 24-hour dependence is local.

transformed ozone time series and use the residual time series as data. An example of a ozone time series in its original and transformed form is shown in Figure 2.5.

We estimate the model by using the two-step ℓ_1-penalized least squares method, using stability selection (Meinshausen and Bühlmann, 2010) to sidestep the problem that the choice of the regularization parameters λ_1 and λ_2 in the first and second step of the two-step ℓ_1-penalized estimation method depends on the unknown values of $\boldsymbol{\beta}^\star$ and $\boldsymbol{\Sigma}$. In Step 1, we include pollution monitors in the 15 most polluted cities in the U.S.A. in 2015—according to the website of the American Lung Association—with probability $.99$ and other pollution monitors with probability $.01$. We excluded 91 pollution monitors in sparsely monitored regions and regions with known omitted monitors—omitted due to a large fraction of missing data—from the sample out of the concern that such monitors may give rise to spurious edges. Most of those monitors are located in sparsely populated and mountainous regions in the Midwest and West.

We compare the two-step ℓ_1-penalized least squares method with an estimate $\widehat{\rho}$ of ρ to the two-step ℓ_1-penalized least squares method with an upper bound on ρ given by $\rho = 250$ and the ℓ_1-penalized least squares method of Basu and Michailidis (2015). The upper bound $\rho = 250$ is based

on scientific evidence (Rao et al., 1997), which suggests that $\rho \leq 250$. The ℓ_1-penalized least squares method of Basu and Michailidis (2015) is the most attractive high-dimensional method available, as discussed in Remark 3 in Section 2.4.

The two-step ℓ_1-penalized least squares method estimates ρ by $\widehat{\rho} = 239$. It is more than 8 times faster than the ℓ_1-penalized least squares method and reduces the out-of-sample 24-hour ahead forecast mean squared error by 4%. If the upper bound $\rho = 250$ is used and hence ρ is not estimated, the two-step ℓ_1-penalized least squares method is more than 18 times faster than the ℓ_1-penalized least squares method.

The graphs estimated by the ℓ_1-penalized least squares method and the two-step ℓ_1-penalized least squares method with estimate $\widehat{\rho} = 239$ and upper bound $\rho = 250$ are shown in Figures 2.1 and 2.6, respectively. It is striking that the ℓ_1-penalized least squares method reports a number of long-distance edges—some of them between monitors separated by more than 2,166 miles. The long-distance edges conflict with scientific evidence, which suggests that dependence local and that $\rho \leq 250$ (e.g., Rao et al., 1997): it is not believed that today's ozone levels on the East Coast can directly affect tomorrow's ozone levels on the West Coast, because ozone cannot travel long distances (see, e.g., Rao et al., 1997). In contrast, the two-step ℓ_1-penalized least squares method reports that the estimated range of 24-hour dependence is $\widehat{\rho} = 239$, which is consistent with scientific evidence (e.g., Rao et al., 1997).

2.6.2 Zooming in: pollution in the Gulf region

We zoom in on the Gulf of Mexico region and consider the 24-hour transport of 6 pollutants: ozone (O_3), particle matter ($PM10$ and $PM2.5$), carbon monoxide (CO), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). The data set consists of $N = 1{,}826$ observations of the 6 pollutants recorded by $k = 199$ monitors between January 2010 and December 2014. 45.2% of the time series are ozone time series, 22.6% are NO_2, 15.6% are $PM25$, 9.1% are SO_2, 5.5% are CO, and 2.0% are $PM10$.

We estimate the model by the two-step ℓ_1-penalized least squares method. In Step 1, we ensure

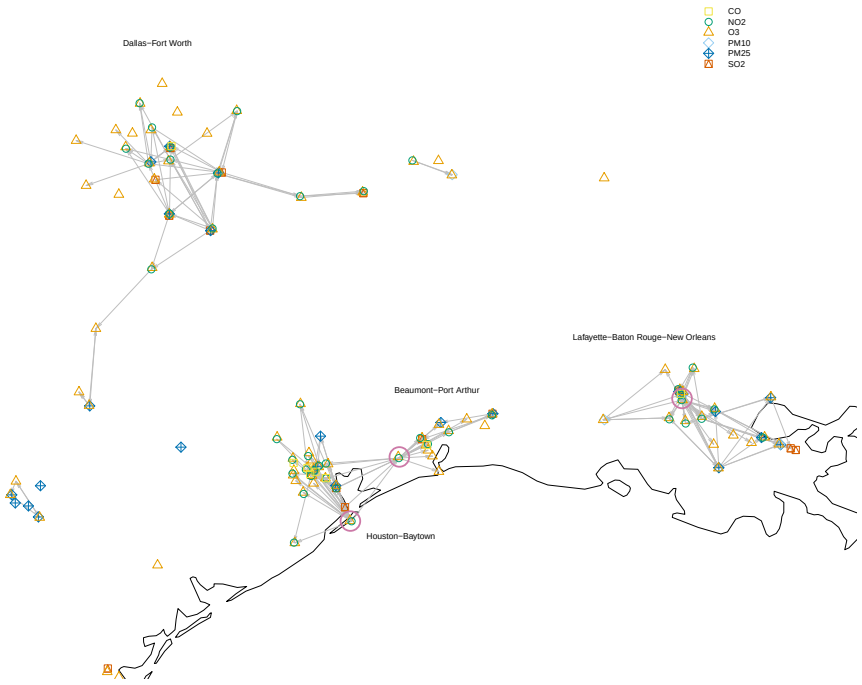


Figure 2.7 : Air pollution in the Gulf of Mexico region: autoregressive coefficients estimated by the two-step ℓ_1-penalized least squares method from daily measurements of 6 pollutants. Monitors are connected by edges if the estimates of the corresponding autoregressive coefficients are non-zero. Monitors with at least 18 outgoing edges are indicated by circles.

that monitors of all 6 pollutants are well-represented in the sample by generating a stratified sample of size 20, where the sample size of monitors of a pollutant is proportional to the total number of monitors of the pollutant in the Gulf of Mexico region. The graph estimated by the two-step ℓ_1-penalized least squares method is presented in Figure 2.7. Most edges are $NO_2 \to NO_2$ edges, while most cross-pollutant edges are $NO_2 \to O_3$ edges, which may be due to the chemical reaction that transforms nitrogen oxides into ozone in the presence of sunlight.

There are two eye-catching facts in Figure 2.7. First, there are 4 clusters, Dallas—Fort Worth, Houston—Baytown, Beaumont—Port Arthur, and Lafayette—Baton Rouge—New Orleans, corresponding to industrial and metropolitan areas in the Gulf of Mexico region. Second, while the dependence structure is sparse and the median number of outgoing edges of monitors is 1, there are 3 monitors with at least 18 outgoing edges, most of which are positive. The large number

of positive outgoing edges—i.e., positive autoregressive coefficients—suggests that pollution at those 3 locations tends to drive up pollution in neighboring regions. It turns out that all of them are home to large industrial complexes, including some of the largest oil refineries in the U.S.A. These findings suggest that neighboring regions have reason to be concerned with the activities of the industrial sectors in those areas.

2.7 Appendix: Proofs of Chapter 2

We prove Theorems 2.1 and 2.2 in Appendices 2.7.1 and 2.7.2, respectively.

2.7.1 Proof of Theorem 2.1

Let $\delta \geq 0$. It is convenient to express the estimator $\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}$ of $\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}$ obtained in Step 2 of the two-step ℓ_1-penalized least squares method as the solution of the M-estimation problem

$$\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} \in \underset{\boldsymbol{\beta}_{[0,\rho-\delta]}}{\arg\min} \left[-2\, \boldsymbol{\beta}_{[0,\rho-\delta]}^{\top}\, \widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} + \boldsymbol{\beta}_{[0,\rho-\delta]}^{\top}\, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\, \boldsymbol{\beta}_{[0,\rho-\delta]} + \lambda_2 \left\| \boldsymbol{\beta}_{[0,\rho-\delta]} \right\|_1 \right].$$

We need three lemmas to prove Theorem 2.1.

Lemma 2.1 *Assume $N \geq c_0\, s \log p$ ($c_0 > 1$). Then, for all $\delta \geq 0$, with at least probability $1 - 2\, \exp(-c_1\, N)$,*

$$\boldsymbol{b}^{\top}\, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\, \boldsymbol{b} \;\; \geq \;\; \alpha\, \|\boldsymbol{b}\|_2^2 - \tau\, \|\boldsymbol{b}\|_1^2 \;\; \textit{for all} \;\; \boldsymbol{b} \in \mathbb{R}^{p(0,\rho-\delta)}. \tag{2.9}$$

Proof. Observe that $\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}$ can be written as $\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]} = \boldsymbol{E}^{\top}\, \widehat{\boldsymbol{\Gamma}}\, \boldsymbol{E}$, where $\boldsymbol{E}$ is a 0-1 elimination matrix of suitable order that eliminates the elements of $\widehat{\boldsymbol{\Gamma}}$ that are not elements of $\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}$. By Assumption 1, for all $\boldsymbol{b} \in \mathbb{R}^{p(0,\rho-\delta)}$,

$$\boldsymbol{b}^{\top}\, \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\, \boldsymbol{b} \;=\; (\boldsymbol{E}\, \boldsymbol{b})^{\top}\, \widehat{\boldsymbol{\Gamma}}\, (\boldsymbol{E}\, \boldsymbol{b}) \;\geq\; \alpha\, \|\boldsymbol{E}\, \boldsymbol{b}\|_2^2 - \tau\, \|\boldsymbol{E}\, \boldsymbol{b}\|_1^2 \;=\; \alpha\, \|\boldsymbol{b}\|_2^2 - \tau\, \|\boldsymbol{b}\|_1^2, \tag{2.10}$$

where $||\boldsymbol{E}\, \boldsymbol{b}||_i = ||\boldsymbol{b}||_i$, $i = 1, 2$, because the p-vector $\boldsymbol{E}\, \boldsymbol{b}$ consists of the $p(0, \rho - \delta)$ elements of $\boldsymbol{b}$ and $p - p(0, \rho - \delta)$ 0's. The lower bound (2.10) holds as long as Assumption 1 holds. By

Proposition 4.2 of Basu and Michailidis (2015), the probability that Assumption 1 is violated is bounded above by $2\,\exp(-c_1\,N)$ provided $N \geq c_0\,s\log p$ ($c_0 > 1$).

Lemma 2.2 *Assume* $N \geq \log p(0, \rho - \delta)$. *Then, for all* $\delta \geq 0$, *with at least probability* $1 - 6\,\exp\left(-c_2 \log p(0, \rho - \delta)\right)$,

$$\|\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_{\infty} \;\;\leq\;\; \mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma})\sqrt{\frac{\log p(0, \rho - \delta)}{N}}. \tag{2.11}$$

Proof. The proof proceeds along the lines of Proposition 4.3 of Basu and Michailidis (2015, supplement, pp. 6–7) by applying concentration inequality (2.11) of Basu and Michailidis (2015) to bound the probability of

$$\|\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_{\infty} \;\;>\;\; 2\,\pi\,\frac{\mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma})}{a}\,\eta,$$

where $a > 0$ and $\eta > 0$. Choosing $\eta = (a/(2\,\pi))\,\sqrt{\log p(0, \rho - \delta)/N}$ gives

$$\|\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_{\infty} \;\;>\;\; \mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma})\sqrt{\frac{\log p(0, \rho - \delta)}{N}}. \tag{2.12}$$

The concentration inequality (2.11) of Basu and Michailidis (2015) and a union bound show that, provided $N \geq \log p(0, \rho - \delta)$, the probability of (2.12) is bounded above by

$$6\,\exp\left(-c\,N\,\min\left(\eta, \eta^2\right) + \log p(0, \rho - \delta)\right) \;\;\leq\;\; 6\,\exp\left(-c_2 \log p(0, \rho - \delta)\right).$$

Lemma 2.3 *Assume that conditions* (2.9) *and* (2.11) *are satisfied and* $\lambda_2 \geq 4\,\mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma})\,\sqrt{\log p(0, \rho - \delta)/N}$. *Then, for all* $\delta \geq 0$,

$$\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_2 \;\;\leq\;\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha}.$$

Proof. By definition of $\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}$, for all $\boldsymbol{\beta}_{[0,\rho-\delta]} \in \mathbb{R}^{p(0,\rho-\delta)}$,

$$\begin{aligned} & -2\,\widehat{\boldsymbol{\beta}}^{\top}_{[0,\rho-\delta]}\,\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} + \widehat{\boldsymbol{\beta}}^{\top}_{[0,\rho-\delta]}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} + \lambda_2\,\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]}\|_1 \\ \leq\;\; & -2\,\boldsymbol{\beta}^{\top}_{[0,\rho-\delta]}\,\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} + \boldsymbol{\beta}^{\top}_{[0,\rho-\delta]}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}_{[0,\rho-\delta]} + \lambda_2\,\|\boldsymbol{\beta}_{[0,\rho-\delta]}\|_1. \end{aligned} \tag{2.13}$$

Set $\boldsymbol{\beta}_{[0,\rho-\delta]} = \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}$ and $\boldsymbol{v} = \widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}$. Then (2.13) reduces to

$$\begin{aligned} &\boldsymbol{v}^{\top}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{v} \\ &\leq 2\,\boldsymbol{v}^{\top}(\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}) + \lambda_2\,(\|\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_1 - \|\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]} - \boldsymbol{v}\|_1). \end{aligned} \tag{2.14}$$

The first term on the right-hand side of (2.14) can be bounded by using condition (2.11) and $\lambda_2 \geq 4\,\mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma})\,\sqrt{\log p(0,\rho-\delta)/N}$:

$$2\,\boldsymbol{v}^{\top}(\widehat{\boldsymbol{\gamma}}_{[0,\rho-\delta]} - \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}) \;\leq\; \frac{\lambda_2}{2}\,\|\boldsymbol{v}\|_1 \;=\; \frac{\lambda_2}{2}\,(\|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 + \|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1), \tag{2.15}$$

where $\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}$ and $\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}$ are the subvectors of $\boldsymbol{v}$ corresponding to the support $\mathbb{S}[0,\rho-\delta]$ of $\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}$ and its complement $\overline{\mathbb{S}}[0,\rho-\delta]$, respectively. The second term on the right-hand side of (2.14) can be bounded as follows:

$$\lambda_2\,(\|\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_1 - \|\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]} - \boldsymbol{v}\|_1) \;\;\leq\;\; \lambda_2\,(\|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 - \|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1) \tag{2.16}$$

using the triangle inequality

$$\|\boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_1 \;\;=\;\; \|\boldsymbol{\beta}^{\star}_{\mathbb{S}[0,\rho-\delta]}\|_1 \;\;\leq\;\; \|\boldsymbol{\beta}^{\star}_{\mathbb{S}[0,\rho-\delta]} - \widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 + \|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1.$$

Therefore, combining (2.14) with (2.15) and (2.16),

$$0 \;\;\leq\;\; \boldsymbol{v}^{\top}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{v} \;\;\leq\;\; \frac{3\,\lambda_2}{2}\|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 - \frac{\lambda_2}{2}\,\|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1. \tag{2.17}$$

Thus, $\|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1 \leq 3\,\|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1$, implying

$$\|\boldsymbol{v}\|_1 \;\;=\;\; \|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 + \|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1 \;\;\leq\;\; 4\,\|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 \;\;\leq\;\; 4\,\sqrt{s}\,\|\boldsymbol{v}\|_2. \tag{2.18}$$

An upper bound on $\boldsymbol{v}^{\top}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{v}$ can therefore be obtained by using (2.17) and (2.18):

$$\boldsymbol{v}^{\top}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{v} \;\;\leq\;\; \frac{3\,\lambda_2}{2}\|\widehat{\boldsymbol{v}}_{\mathbb{S}[0,\rho-\delta]}\|_1 - \frac{\lambda_2}{2}\,\|\widehat{\boldsymbol{v}}_{\overline{\mathbb{S}}[0,\rho-\delta]}\|_1 \;\;\leq\;\; 2\,\lambda_2\,\|\boldsymbol{v}\|_1,$$

implying

$$\frac{1}{2}\,\boldsymbol{v}^{\top}\,\widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\,\boldsymbol{v} \;\;\leq\;\; \lambda_2\,\|\boldsymbol{v}\|_1 \;\;\leq\;\; 4\,\sqrt{s}\,\lambda_2\,\|\boldsymbol{v}\|_2.$$

A lower bound on $\boldsymbol{v}^\top \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\, \boldsymbol{v}$ can be derived by using Lemma 2.1 and (2.18) along with $s\,\tau \leq \alpha/32$, giving

$$\boldsymbol{v}^\top \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\, \boldsymbol{v} \;\geq\; \alpha\,\|\boldsymbol{v}\|_2^2 - \tau\,\|\boldsymbol{v}\|_1^2 \;\geq\; \alpha\,\|\boldsymbol{v}\|_2^2 - \tau\,16\,s\,\|\boldsymbol{v}\|_2^2 \;\geq\; \frac{\alpha}{2}\,\|\boldsymbol{v}\|_2^2.$$

Combining the upper and lower bounds on $\boldsymbol{v}^\top \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\, \boldsymbol{v}$ gives

$$\frac{\alpha}{4}\,\|\boldsymbol{v}\|_2^2 \;\;\leq\;\; \frac{1}{2}\,\boldsymbol{v}^\top \widehat{\boldsymbol{\Gamma}}_{[0,\rho-\delta],[0,\rho-\delta]}\, \boldsymbol{v} \;\;\leq\;\; 4\,\sqrt{s}\,\lambda_2\,\|\boldsymbol{v}\|_2,$$

implying

$$\|\boldsymbol{v}\|_2 \;\;=\;\; \|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^\star_{[0,\rho-\delta]}\|_2 \;\;\leq\;\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha}.$$

Proof. Theorem 2.1. By Lemma 2.3 with $\delta = 0$, as long as conditions (2.9) and (2.11) are satisfied,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \;\;=\;\; \|\widehat{\boldsymbol{\beta}}_{[0,\rho]} - \boldsymbol{\beta}^\star_{[0,\rho]}\|_2 \;\;\leq\;\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha}, \tag{2.19}$$

where we used the fact that all elements of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^\star$ corresponding to edges at distances $d > \rho$ are 0. The upper bound (2.19) holds as long as conditions (2.9) and (2.11) hold. By Lemmas 2.1 and 2.2 with $\delta = 0$ along with $N \geq c_0\,s\log p \geq \log p(0,\rho)$ ($c_0 > 1$) and a union bound, the probability that (2.9) or (2.11) are violated is bounded above by

$$2\,\exp(-c_1\,N) + 6\,\exp(-c_2 \log p(0,\rho)).$$

2.7.2 Proof of Theorem 2.2

We need three additional lemmas to prove Theorem 2.2.

Lemma 2.4 *For all $\delta > 0$, the probability that none of the nodes $i \in \mathbb{S}(\delta)$ is sampled is bounded above by*

$$\exp\left(-\sum_{i\in\mathbb{S}(\delta)} \theta_i\right).$$

Proof. By definition of $\rho > 0$, for all $\delta > 0$, there exists at least one node with incoming edges at distances $d \in [\rho - \delta, \rho]$, thus $\mathbb{S}(\delta)$ is non-empty. Since nodes i are sampled independently with probabilities $0 < \theta_i < 1$, the probability that none of the nodes $i \in \mathbb{S}(\delta)$ is sampled is bounded above by

$$\exp\left(\sum_{i \in \mathbb{S}(\delta)} \log(1 - \theta_i)\right) \;\; \leq \;\; \exp\left(-\sum_{i \in \mathbb{S}(\delta)} \theta_i\right).$$

Lemma 2.5 *Let $\beta^\star_{\min} \geq 32\,\sqrt{s}\,\lambda_1/\alpha$, where $\lambda_1 \geq 4\,\mathbb{Q}(\boldsymbol{\beta}^\star, \boldsymbol{\Sigma})\,\sqrt{\log p/N}$. Then, for any $\delta > 0$ and any non-empty subset $\mathcal{A} \subseteq \mathbb{S}(\delta)$, the probability that none of the incoming edges of nodes $i \in \mathcal{A}$ at distances $d \in [\rho - \delta, \rho]$ is detected is bounded above by*

$$2\,\exp(-c_1\,N) + 6\,\exp(-c_2 \log p).$$

Proof. By definition of $\rho > 0$, for all $\delta > 0$, there exists at least one node with incoming edges at distances $d \in [\rho - \delta, \rho]$, thus $\mathbb{S}(\delta)$ is non-empty. Consider any non-empty subset $\mathcal{A} \subseteq \mathbb{S}(\delta)$. Let $\mathbb{G}$ be the event that all incoming edges of all nodes $i \in \mathcal{A}$ are detected and $\mathcal{B}$ be its complement. Then the event that none of the incoming edges of nodes $i \in \mathcal{A}$ at distances $d \in [\rho - \delta, \rho]$ is detected is contained in $\mathcal{B}$ and the probability of the event of interest is bounded above by the probability of $\mathcal{B}$. To bound the probability of $\mathcal{B}$, let $\widehat{\boldsymbol{\beta}}_{\mathcal{N}}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ be solutions of optimization problems (2.1) and (2.2), respectively, and observe that $\mathbb{G}$ is implied by

$$\frac{2}{\beta^\star_{\min}}\,\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}}\|_\infty \;\; \leq \;\; 1.$$

Since

$$\frac{2}{\beta^\star_{\min}}\,\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}}\|_\infty \;\; \leq \;\; \frac{2}{\beta^\star_{\min}}\,\|\widehat{\boldsymbol{\beta}}_{\mathcal{N}} - \boldsymbol{\beta}^\star_{\mathcal{N}}\|_\infty,$$

we have, by Assumptions 1 and 2 and $\beta^\star_{\min} \geq 32\,\sqrt{s}\,\lambda_1/\alpha$,

$$\frac{2}{\beta^\star_{\min}}\,\|\widehat{\boldsymbol{\beta}}_{\mathcal{N}} - \boldsymbol{\beta}^\star_{\mathcal{N}}\|_\infty \;\; \leq \;\; \frac{2}{\beta^\star_{\min}}\,\|\widehat{\boldsymbol{\beta}}_{\mathcal{N}} - \boldsymbol{\beta}^\star_{\mathcal{N}}\|_2 \;\; \leq \;\; \frac{2}{\beta^\star_{\min}}\,\frac{16\,\sqrt{s}\,\lambda_1}{\alpha} \;\; \leq \;\; 1. \tag{2.20}$$

The bound $||\widehat{\boldsymbol{\beta}}_{\mathcal{N}} - \boldsymbol{\beta}^\star_{\mathcal{N}}||_2 \leq 16\,\sqrt{s}\,\lambda_1/\alpha$ used in (2.20) follows from Proposition 4.1 of Basu and Michailidis (2015) and holds as long as Assumptions 1 and 2 hold. Therefore, $\mathbb{G}$ occurs as long

as both Assumptions 1 and 2 hold, whereas $\mathcal{B}$ occurs when either Assumption 1 or Assumption 2 or both are violated. A union bound along with $N \geq c_0\, s \log p \geq \log p$ ($c_0 > 1$) shows that the probability of $\mathcal{B}$, and thus the event of interest, is bounded above by

$$2\, \exp(-c_1\, N) + 6\, \exp(-c_2 \log p), \tag{2.21}$$

where the two terms in (2.21) are upper bounds on the probabilities that Assumption 1 or Assumption 2 are violated, which follow from Propositions 4.2 and 4.3 of Basu and Michailidis (2015), respectively.

Lemma 2.6 *Consider $N \geq c_0\, s \log p$ ($c_0 > 1$) observations from a stable L-th order vector autoregressive process with radius $\rho > 0$. Assume that components i are sampled independently with probabilities $0 < \theta_i < 1$, the minimum signal strength is $\beta^{\star}_{\min} = \min_{i \in \mathbb{S}} |\beta^{\star}_i| \geq 32\, \sqrt{s}\, \lambda_1 / \alpha > 0$, and the regularization parameter λ_1 satisfies*

$$\lambda_1 \;\; \geq \;\; 4\, \mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma}) \sqrt{\frac{\log p}{N}}.$$

Then, for all $\delta > 0$,

$$\mathbb{P}\left(\widehat{\rho} \, - \rho \, < \, -\delta\right) \;\; \leq \;\; 2\, \exp(-c_1\, N) + 6\, \exp(-c_2 \log p) + \exp\left(- \sum_{i \in \mathbb{S}(\delta)} \theta_i \right).$$

Proof. By definition of $\rho > 0$, for all $\delta > 0$, there exists at least one node with incoming edges at distances $d \in [\rho - \delta, \rho]$, thus $\mathbb{S}(\delta)$ is non-empty. Let $\mathbb{G}_1$ be the event that at least one node $i \in \mathbb{S}(\delta)$ with incoming edges at distances $d \in [\rho - \delta, \rho]$ is sampled and that at least one of its incoming edges at distances $d \in [\rho - \delta, \rho]$ is detected and $\mathbb{G}_2$ be the event that at least one false-positive incoming edge of nodes $i \in \mathbb{S}$ at distances $d \in [\rho - \delta, \infty)$ is reported. Then the event $\widehat{\rho} \, - \rho \geq -\delta$ is equivalent to the event $\mathbb{G}_1 \cup \mathbb{G}_2$. Thus, the probability of event $\widehat{\rho} \, - \rho \geq -\delta$ is bounded below by

$$\begin{aligned} \mathbb{P}(\widehat{\rho} \, - \rho \geq -\delta) \;\; &= \;\; \mathbb{P}(\mathbb{G}_1 \cup \mathbb{G}_2) \;\geq\; \mathbb{P}(\mathbb{G}_1) \\ &\geq \;\; 1 - 2\, \exp(-c_1\, N) - 6\, \exp(-c_2 \log p) - \exp\left(- \sum_{i \in \mathbb{S}(\delta)} \theta_i \right), \end{aligned}$$

where we used a union bound along with Lemmas 2.4 and 2.5 to bound the probability of the complement of event $\mathbb{G}_1$.

Proof. Theorem 2.2. For all $\delta > 0$,

$$\begin{aligned} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_2 \, > \, \frac{16\,\sqrt{s}\,\lambda_2}{\alpha} \right) \;\; &\leq \;\; \mathbb{P}\left(\widehat{\rho} \, - \rho \, < \, -\delta\right) \\ + \;\; \mathbb{P}\left(\left(\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_2 \, > \, \frac{16\,\sqrt{s}\,\lambda_2}{\alpha} \right) \cap \left(\widehat{\rho} \, - \rho \geq -\delta \right) \right). \end{aligned} \tag{2.22}$$

We bound the two terms on the right-hand side of (2.22) one by one.

First term on the right-hand side of (2.22). By Lemma 2.6, the first term on the right-hand side of (2.22) is bounded above by

$$\mathbb{P}\left(\widehat{\rho} \, - \rho \, < \, -\delta\right) \;\; \leq \;\; 2\,\exp(-c_1\,N) + 6\,\exp(-c_2 \log p) + \exp\left(- \sum_{i \in \mathbb{S}(\delta)} \theta_i \right). \tag{2.23}$$

Second term on the right-hand side of (2.22). We are interested in the intersection of the event that $\widehat{\rho} \, - \rho \geq -\delta$ and the event that

$$\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_2 \;\; > \;\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha}, \tag{2.24}$$

where $\lambda_2 \geq 4\,\mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma})\,\sqrt{\log p(0, \rho - \delta)/N}$. By Lemma 2.3 and $\lambda_2 \geq 4\,\mathbb{Q}(\boldsymbol{\beta}^{\star}, \boldsymbol{\Sigma})\,\sqrt{\log p(0, \rho - \delta)/N}$, as long as conditions (2.9) and (2.11) are satisfied, $\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_2$ is bounded above by

$$\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_2 \;\; \leq \;\; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha}. \tag{2.25}$$

By Lemmas 2.1 and 2.2 along with $N \geq c_0\,s \log p \geq \log p(0, \rho - \delta)$ ($c_0 > 1$) and a union bound, the probability that (2.9) or (2.11) are violated is bounded above by

$$2\,\exp(-c_1\,N) + 6\,\exp(-c_2 \log p(0, \rho - \delta)). \tag{2.26}$$

Thus, the second term on the right-hand side of (2.22) is bounded above by (2.26).

Conclusion. Combining (2.22), (2.23), and (2.26) shows that

$$\begin{aligned}
&\mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}_{[0,\rho-\delta]} - \boldsymbol{\beta}^{\star}_{[0,\rho-\delta]}\|_2 \; > \; \frac{16\,\sqrt{s}\,\lambda_2}{\alpha} \right) \\
\leq \;& 2\,\exp(-c_1\,N) + 6\,\exp(-c_2 \log p) + \exp\left(- \sum_{i \in \mathbb{S}(\delta)} \theta_i \right) \\
+ \;& 2\,\exp(-c_1\,N) + 6\,\exp(-c_2 \log p(0, \rho - \delta)) \\
\leq \;& 4\,\exp(-c_1\,N) + 12\,\exp(-c_2 \log p(0, \rho - \delta)) + \exp\left(- \sum_{i \in \mathbb{S}(\delta)} \theta_i \right),
\end{aligned} \tag{2.27}$$

where we used the fact that $p(0, \rho - \delta) \leq p$ for all $\delta > 0$.

Chapter 3

Massive-scale estimation of exponential-family random graph models with additional structure

Abstract

Chapter 3 considers exponential-family random graph models for modeling complex dependencies in network data. Similar to Chapter 2, additional structure is often available: e.g., it is known that many networks, such as insurgencies and terrorist networks, are local in nature. Such a local structure decomposes random graphs into independent subgraphs with local dependence, which enables development of scalable methods. However, in contrast to the setting of Chapter 2, such additional structure is usually unobserved and has to be estimated. Hence, a two-step likelihood-based approach exploiting additional structure is proposed. The first step estimates the local structure underlying the random graph. The second step estimates the model parameters given the estimated local structure of the random graph. Both steps can be implemented in parallel, which enables massive-scale estimation. Theoretical justification is provided for the two-step likelihood-based approach and its advantages are demonstrated by simulations and an application to a large Amazon product network.

3.1 Introduction

Models of network data are in high demand in statistics and related areas (Kolaczyk, 2009). Such models are useful for studying insurgent and terrorist networks, contact networks facilitating the

---

The contents of Chapter 3 have been submitted to a peer-reviewed journal: Babkin, S. and M. Schweinberger. Massive-scale estimation of exponential-family random graph models with local dependence

spread of infectious diseases, social networks, the World Wide Web, and other networks of interest.

A flexible approach to modeling network data is based on exponential-family random graph models (Frank and Strauss, 1986; Lusher et al., 2013). While exponential-family random graph models are widely used by network scientists (Lusher et al., 2013), statistical inference for exponential-family random graph models is challenging. One reason is that some models, such as the classic models of Frank and Strauss (1986), are ill-posed and allow edges to depend on many other edges in the network. Applying such models to large networks is problematic: e.g., a friendship between two users of Facebook may depend on friendships with other users, but it is not plausible that it depends on friendships with billions of other users. As a result, such models can induce strong dependence among edges, which in turn can lead to model degeneracy (Handcock, 2003; Schweinberger, 2011; Chatterjee and Diaconis, 2013). Model degeneracy means that models place much probability mass on sufficient statistics close to the boundary of the convex hull of the sufficient statistics (Schweinberger, 2011). If graphs are generated by such models, the sufficient statistics of the generated graphs tend to be close to the boundary of the convex hull, which implies that maximum likelihood estimators either do not exist at all or are hard to obtain by maximum likelihood algorithms (Handcock, 2003; Rinaldo et al., 2009). In addition, the results of Shalizi and Rinaldo (2013) suggest that maximum likelihood estimators of some models may not be consistent.

To address the problems of exponential-family random graph models, Schweinberger and Handcock (2015) proposed exponential-family random graph models with additional structure. The basic idea is that random graphs are endowed with additional structure in the form of neighborhood structure and that the dependence induced by the models is local in the sense that it is restricted to neighborhoods. Such exponential-family random graph models with additional structure, which we call exponential-family random graph models with local dependence, have at least two important advantages over exponential-family random graph models without additional structure. First, local dependence induces weak dependence and models with weak dependence are less prone to model degeneracy (e.g., Schweinberger and Stewart, 2016, Corollary 1). Second, models with local dependence satisfy a weak form of self-consistency in the sense that these models are

consistent under neighborhood sampling (Schweinberger and Handcock, 2015, Theorem 1). While the notion of consistency under neighborhood sampling is weaker than the notion of consistency under sampling of Shalizi and Rinaldo (2013), it enables consistent estimation of neighborhood-dependent parameters. Schweinberger and Stewart (2016) showed that when the neighborhood structure is known and the neighborhoods grow at the same rate, M-estimators of neighborhood-dependent parameters of canonical and curved exponential-family random graph models with local dependence are consistent. Schweinberger (2017) showed that when the neighborhood structure is unknown, it can be recovered with high probability.

While these consistency results suggest that exponential-family random graph models with additional structure have important conceptual and statistical advantages over exponential-family random graph models without additional structure, there are no methods for estimating them from large random graphs. Schweinberger and Handcock (2015) used Bayesian methods to estimate them, but Bayesian methods are too time-consuming to be applied to random graphs with more than one hundred nodes. We pave the ground for massive-scale estimation of such models by proposing a two-step likelihood-based approach that exploits model structure for the purpose of parallel computing. The main idea is that random graphs can be decomposed into subgraphs with local dependence and hence parallel computing can be used to compute the contributions of subgraphs to the likelihood function. Motivated by these considerations, we propose a two-step likelihood-based approach. The first step estimates the neighborhood structure and decomposes random graphs into subgraphs with local dependence. The decomposition of the random graph relies on approximations of the likelihood function that are supported by theoretical results. The second step estimates parameters given the estimated neighborhood structure by using Monte Carlo maximum likelihood methods (Hunter and Handcock, 2006). Both steps can be implemented in parallel, which enables massive-scale estimation on multi-core computers or computing clusters. We demonstrate the advantages of the two-step likelihood-based approach by simulations and an application to a large Amazon product network.

This chapter is structured as follows. Section 3.2 introduces models. Section 3.3 discusses

likelihood-based inference based on approximations of the likelihood function that are supported by theoretical results and Section 3.4 takes advantage of such approximations to estimate models. Section 3.5 presents simulation results and Section 3.6 applications.

Other, related literature. It is worth noting that a recent paper by Thiemichen and Kauermann (2017) considers estimating nonparametric exponential-family random graph models from large networks. We consider a more challenging task than Thiemichen and Kauermann (2017), because we consider models with additional structure in the form of neighborhood structure and focus on the estimation of neighborhood structure as well as parameters, whereas Thiemichen and Kauermann (2017) consider models without neighborhood structure and focus on the estimation of parameters.

3.2 Models

To introduce exponential-family random graph models with additional structure, let $\mathcal{A} = \{1, \ldots, n\}$ be a set of nodes and $\mathcal{E} \subseteq \mathcal{A} \times \mathcal{A}$ be a subset of edges between pairs of nodes. Throughout, we consider undirected random graphs without self-edges, i.e., we assume that $(i, i) \notin \mathcal{E}$ and $(i, j) \in \mathcal{E}$ implies and is implied by $(j, i) \in \mathcal{E}$, but all models discussed here can be extended to directed random graphs. We regard edges as random variables denoted by $X_{i,j}$, where $X_{i,j}$ takes on values in a countable set $\mathbb{X}_{i,j}$, i.e., we consider both binary and non-binary, network count data. We write $\boldsymbol{X} = (X_{i,j})_{i<j}^{n}$ and $\mathbb{X} = \bigtimes_{i<j}^{n} \mathbb{X}_{i,j}$.

We assume that the random graph $\boldsymbol{X}$ is governed by an exponential family with countable support $\mathbb{X}$ and probability mass functions of the form

$$p_{\boldsymbol{\eta}}(\boldsymbol{x}) \;=\; \exp(\langle \boldsymbol{\eta},\, s(\boldsymbol{x})\rangle - \psi(\boldsymbol{\eta})), \qquad \boldsymbol{x} \in \mathbb{X}, \tag{3.1}$$

where $\langle \boldsymbol{\eta},\, s(\boldsymbol{x})\rangle$ is the inner product of a natural parameter vector $\boldsymbol{\eta} \in \mathbb{N} = \{\boldsymbol{\eta} \in \mathbb{R}^{\dim(\boldsymbol{\eta})} : \psi(\boldsymbol{\eta}) < \infty\}$ and a vector of sufficient statistics $s : \mathbb{X} \mapsto \mathbb{R}^{\dim(\boldsymbol{\eta})}$ and

$$\psi(\boldsymbol{\eta}) \;=\; \log \sum_{\boldsymbol{x}' \in \mathbb{X}} \exp\left(\langle \boldsymbol{\eta},\, s(\boldsymbol{x}')\rangle\right).$$

We consider here exponential-family random graph models with additional structure, which have important conceptual and statistical advantages over exponential-family random graph models without additional structure, as discussed in Section 3.1. The additional structure takes the form of a partition of the set of nodes $\mathcal{A}$ into subsets of nodes $\mathcal{A}_1, \ldots, \mathcal{A}_K$, called neighborhoods, such that the dependence is local. To introduce the notion of local dependence, assume that $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \boldsymbol{\Xi} \subseteq \mathbb{N}$ is a function of a neighborhood-dependent parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and a neighborhood membership vector $\boldsymbol{z} \in \mathbb{Z}$, where $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ consists of neighborhood memberships $\boldsymbol{z}_i$ such that $z_{i,k} = 1$ if node i belongs to neighborhood $\mathcal{A}_k$ and $z_{i,k} = 0$ otherwise. We henceforth denote by $\boldsymbol{X}_{k,k} = (X_{i,j})^n_{i<j:\, z_{i,k}=z_{j,k}=1}$ the sequence of within-neighborhood edge variables of neighborhood $\mathcal{A}_k$ ($k = 1, \ldots, K$).

Definition. Local dependence. An exponential-family random graph model with countable support $\mathbb{X}$ satisfies local dependence as long as

$$p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) \;=\; \prod_{k=1}^{K} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k}) \prod_{l=1}^{k-1} \prod_{i,j:\, z_{i,k}=1,\, z_{j,l}=1}^{n} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(x_{i,j}), \qquad \boldsymbol{x} \in \mathbb{X}. \tag{3.2}$$

In other words, the dependence is local in the sense that it is confined to within-neighborhood subgraphs. Schweinberger and Stewart (2016, Corollary 1) showed that models with local dependence induce weak dependence and are less prone to model degeneracy than models without local dependence as long as the neighborhoods are not too large. Schweinberger and Stewart (2016) and Schweinberger (2017) detail conditions under which consistent estimation of models with local dependence is possible.

An example of exponential-family random graph models with local dependence and support $\mathbb{X} = \{0, 1\}^{\binom{n}{2}}$ is given by

$$p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) \;\propto\; \exp\left(\sum_{k\leq l}^{K} \theta_{1,k,l} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,l} + \sum_{k=1}^{K} \theta_{2,k,k} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,k} \max_{h\neq i,j} x_{i,h}\, x_{j,h}\, z_{h,k} \right).$$

The model includes between- and within-neighborhood edge terms and within-neighborhood transitive edge terms. The transitive edge statistics $x_{i,j}\, z_{i,k}\, z_{j,k} \max_{h\neq i,j} x_{i,h}\, x_{j,h}\, z_{h,k}$ capture transitive closure and induce dependence among edges, but the dependence is confined to within-

neighborhood subgraphs and is hence local.

A special case of exponential-family random graph models with local dependence are stochastic block models (Nowicki and Snijders, 2001): e.g., consider the model above and let $\boldsymbol{\theta}_1 = (\theta_{1,k,l})_{k \leq l}^{K}$ and $\boldsymbol{\theta}_2 = (\theta_{2,k,k})_{k=1}^{K}$; then $\boldsymbol{\theta}_2 = \mathbf{0}$ reduces the model to a model with between- and within-neighborhood edge terms, which corresponds to a stochastic block model that assumes edges to be independent Bernoulli$(\mu_{k,l})$ random variables with $\mu_{k,l} = \text{logit}^{-1}(\theta_{1,k,l})$.

3.3 Likelihood-based inference

While it is natural to base statistical inference concerning $\boldsymbol{z}$ and $\boldsymbol{\theta}$ on the likelihood function, likelihood-based inference for exponential-family random graph models with local dependence is challenging. The main reason is that the probability mass function $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ is intractable, because the within-neighborhood probability mass functions $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k})$ are intractable. The intractability of $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k})$ is rooted in the fact that its normalizing constant is a sum over all possible within-neighborhood subgraphs of neighborhood $\mathcal{A}_k$, which cannot be computed unless $\mathcal{A}_k$ is small, i.e., $|\mathcal{A}_k| \ll 10$ $(k = 1, \ldots, K)$.

To facilitate likelihood-based inference, we introduce tractable approximations of the intractable probability mass function $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ in Section 3.3.1 and support them by theoretical results in Section 3.3.2. A statistical algorithm that takes advantage of such approximations is introduced in Section 3.4.

3.3.1 Approximate likelihood functions: motivation

Suppose that we want to estimate both $\boldsymbol{z}$ and $\boldsymbol{\theta}$. It is natural to estimate them by using an iterative algorithm that cycles through updates of $\boldsymbol{z}$ and $\boldsymbol{\theta}$ as follows:

1. Update $\boldsymbol{z}$ given $\boldsymbol{\theta}$.

2. Update $\boldsymbol{\theta}$ given $\boldsymbol{z}$.

The algorithm sketched above is generic and cannot be used in practice, but regardless of which specific algorithm is used—whether EM, Monte Carlo EM, variational EM, Bayesian Markov chain Monte Carlo, or other algorithms—most of them have in common that Step 1 is either infeasible or time-consuming, whereas Step 2 is less problematic than Step 1.

Step 1 Step 1 is either infeasible or time-consuming, because the probability mass function $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ is intractable. To demonstrate, consider a Bayesian Markov chain Monte Carlo algorithm that updates $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ given $\boldsymbol{\theta}$ by Gibbs sampling. Gibbs sampling of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ turns out to be infeasible, because the full conditional distributions of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ depend on the intractable within-neighborhood probability mass functions $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,k})$. One could approximate them by Monte Carlo samples of within-neighborhood subgraphs, but such approximations may not generate Markov chain Monte Carlo samples from the target distribution (Liang et al., 2016) and are problematic on computational grounds:

- Using Monte Carlo approximations of within-neighborhood probability mass functions is infeasible when the number of nodes n is large, because such approximations are needed for each update of each of the n neighborhood memberships $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$.
- Worse, the n neighborhood memberships $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ cannot be updated in parallel, because the neighborhood membership of one node depends on the neighborhood memberships of other nodes.

Therefore, Step 1 is infeasible when n is large.

Step 2 Step 2 is less problematic than Step 1. While the probability mass function $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ is intractable and may have to be approximated by Monte Carlo methods (Hunter and Handcock, 2006), such Monte Carlo approximations are needed once to update $\boldsymbol{\theta}$ given $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, whereas Monte Carlo approximations are needed n times to update $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ given $\boldsymbol{\theta}$ one by one, which renders Step 1 infeasible when n is large. In addition, the probability mass function $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$

decomposes into between- and within-neighborhood probability mass functions $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}_{k,l})$ and hence within-neighborhood probability mass functions can be approximated in parallel.

Approximations To enable feasible updates of $\boldsymbol{z}$ given $\boldsymbol{\theta}$ when n is large, we are interested in approximating the intractable probability mass function $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ by a tractable probability mass function. To do so, we confine attention to exponential-family random graph models with between- and within-neighborhood edge terms of the form $\sum_{k\leq l}^{K} \theta_{1,k,l} \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,l}$ with parameter vector $\boldsymbol{\theta}_1 = (\theta_{1,k,l})_{k\leq l}^{K}$ and additional model terms with parameter vector $\boldsymbol{\theta}_2$ such that $\boldsymbol{\theta}_2 = \mathbf{0}$ eliminates the additional model terms. An example is given by the edge and transitive edge model in Section 3.2. Such models have two useful properties:

- The probability mass functions $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ impose the same probability law on between-neighborhood subgraphs.

- The probability mass function $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ is tractable, because edges between and within neighborhoods are independent given $\boldsymbol{z}$.

We henceforth approximate $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ by $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$, which corresponds to the probability mass function of a stochastic block model.

The idea underlying the approximation is that when the neighborhoods are not too large, most of the random graph corresponds to between-neigborhood subgraphs. Since $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ impose the same probability law on between-neighborhood subgraphs, $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ agree on most of the random graph. Therefore, $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ can be approximated by $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ for the purpose of updating $\boldsymbol{z}$ given $\boldsymbol{\theta}$. Suppose, e.g., that we consider to update $\boldsymbol{z}$ given $\boldsymbol{\theta}$ by replacing $\boldsymbol{z}$ by some $\boldsymbol{z}' \neq \boldsymbol{z}$. We may decide to do so if the loglikelihood ratio

$$\log \frac{p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}')}(\boldsymbol{x})}{p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})} \quad = \quad \log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$$

is large. If $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ can be approximated by $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$, we can base the decision on $\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ rather than $\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$, because

$$\begin{aligned}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) &= [\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})] \\ &+ [\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z}')}(\boldsymbol{x})] \\ &- [\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})].\end{aligned}$$

Therefore, as long as

$$\max_{\boldsymbol{z}} |\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})|$$

is small, we have

$$\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) \approx \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z}')}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x}).$$

The advantage of approximating $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ by $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ is that there exist methods for stochastic block models to estimate the neighborhood structure from large networks (e.g., Daudin et al., 2008; Rohe et al., 2011; Amini et al., 2013; Vu et al., 2013). We take advantage of such methods in Section 3.4, but we first shed light on the conditions under which $\max_{\boldsymbol{z}} |\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})|$ is small.

3.3.2 Approximate likelihood functions: theoretical results

We show that updates of $\boldsymbol{z}$ given $\boldsymbol{\theta}$ can be based on $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ rather than $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ by showing that

$$\max_{\boldsymbol{z}} |\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{X})|$$

is small with high probability provided that the neighborhoods are not too large and the random graph is not too sparse.

To do so, we make the following assumptions. We assume that $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \boldsymbol{\Xi}$ and that $\boldsymbol{\Xi} \subseteq \text{int}(\mathbb{N})$ is a subset of the interior $\text{int}(\mathbb{N})$ of the natural parameter space $\mathbb{N}$. Let $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{\eta}^\star}$ be the expectation under the data-generating parameter vector $\boldsymbol{\eta}^\star \equiv \boldsymbol{\eta}(\boldsymbol{\theta}^\star, \boldsymbol{z}^\star)$, where $(\boldsymbol{\theta}^\star, \boldsymbol{z}^\star) \in \boldsymbol{\Theta} \times \mathbb{Z}$

denotes the data-generating values of $(\boldsymbol{\theta}, \boldsymbol{z}) \in \boldsymbol{\Theta} \times \mathbb{Z}$. We denote by $d : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}_0^+$ the Hamming metric, which is defined by

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) \;\; = \;\; \sum_{i<j}^{n} \mathbb{1}_{x_{1,i,j} \neq x_{2,i,j}}, \quad (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{X} \times \mathbb{X},$$

where $\mathbb{1}_{x_{1,i,j} \neq x_{2,i,j}}$ is 1 if $x_{1,i,j} \neq x_{2,i,j}$ and is 0 otherwise. This metric calculates the number of edge variables in which random graphs $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ differ. The ℓ_1-, ℓ_2-, and ℓ_∞-norm of vectors are denoted by $\|.\|_1$, $\|.\|_2$, and $\|.\|_\infty$, respectively. In the following, we denote by $n_{\max}(\boldsymbol{z})$ the size of the largest neighborhood under $\boldsymbol{z} \in \mathbb{Z}$. The size of the largest data-generating neighborhood is denoted by $\|\mathcal{A}\|_\infty = \max_{1 \leq k \leq K} |\mathcal{A}_k|$. The main assumptions can then be stated as follows.

[C.1] There exists $c > 0$ and $n_0 > 0$ such that, for all $n > n_0$, all $\boldsymbol{\eta} \in \mathbb{R}^{\dim(\boldsymbol{\eta})}$, and all $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{X} \times \mathbb{X}$,

$$|\langle \boldsymbol{\eta},\, s(\boldsymbol{x}_1) - s(\boldsymbol{x}_2) \rangle| \;\; \leq \;\; c\, d(\boldsymbol{x}_1, \boldsymbol{x}_2)\; n_{\max}(\boldsymbol{z})\; \log n.$$

[C.2] There exists $c > 0$ and $n_0 > 0$ such that, for all $n > n_0$, all $(\boldsymbol{\theta}_{k,l,1},\, \boldsymbol{\theta}_{k,l,2}) \in \boldsymbol{\Theta}_{k,l} \times \boldsymbol{\Theta}_{k,l}$, and all $(\boldsymbol{\theta}, \boldsymbol{z}) \in \boldsymbol{\Theta} \times \mathbb{Z}$,

$$\left| \left\langle \boldsymbol{\eta}_{k,l}(\boldsymbol{\theta}_{k,l,1}, \boldsymbol{z}) - \boldsymbol{\eta}_{k,l}(\boldsymbol{\theta}_{k,l,2}, \boldsymbol{z}),\, \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}\, s_{k,l}(\boldsymbol{X}) \right\rangle \right| \;\; \leq \;\; c\, \|\boldsymbol{\theta}_{k,l,1} - \boldsymbol{\theta}_{k,l,2}\|_2\; n_{\max}(\boldsymbol{z})^2\; \log n,$$

where $\boldsymbol{\eta}_{k,l}(\boldsymbol{\theta}_{k,l}, \boldsymbol{z})$, $\boldsymbol{\theta}_{k,l}$, and $s_{k,l}(\boldsymbol{x})$ denote the subvectors of $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{z})$, $\boldsymbol{\theta}$, and $s(\boldsymbol{x})$ corresponding to the subgraph between neighborhoods k and l ($k < l$) or the subgraph of neighborhood k ($k = l$) and $\boldsymbol{\Theta}_{k,l}$ is a compact subset of $\mathbb{R}^{\dim(\boldsymbol{\theta}_{k,l})}$ ($k \leq l = 1, \ldots, K$).

Conditions [C.1] and [C.2] are satisfied by most exponential-family random graph models of interest, which can be shown by arguments along the lines of Schweinberger (2017, Corollaries 1 and 2). We note that conditions [C.1] and [C.2] cover the size-dependent parameterizations used in Sections 3.5 and 3.6, which give rise to the logarithmic terms in [C.1] and [C.2].

The following result shows that $\max_{\boldsymbol{z}} |\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = \boldsymbol{0}, \boldsymbol{z})}(\boldsymbol{X})|$ is small with high probability provided that the neighborhoods are not too large and the random graph is not too sparse, where the maximum is taken over a subset of well-behaved neighborhood structures.

Theorem 3.1 *Suppose that a random graph is governed by an exponential-family random graph model with countable support $\mathbb{X}$ and local dependence satisfying conditions [C.1] and [C.2]. Let $\mathbb{S} \subseteq \mathbb{Z}$ be a subset of neighborhood structures such that $n_{\max}(\boldsymbol{z}) \leq n_{\max}$ for all $\boldsymbol{z} \in \mathbb{S}$, where $n_{\max}$ may increase as a function of the number of nodes n provided $n_{\max} \leq n$. Then, for all $\delta > 0$, there exist $c > 0$ and $n_0 > 0$ such that, for all $n > n_0$,*

$$\mathbb{P}\left(\max_{\boldsymbol{z} \in \mathbb{S}} |\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X})| \geq c\,(1+\delta)^{1/2}\,\|\mathcal{A}\|_\infty^2\, n_{\max}^2\, n\,(\log n)^{3/2} \right) \leq \epsilon,$$

where $\epsilon = 2\,\exp\left(-\delta\, n \log n\,/\,4\right)$.

The proof of Theorem 3.1 can be found in Section 3.7. The basic idea underlying Theorem 3.1 is that the deviation $\max_{\boldsymbol{z}} |\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})|$ cannot be too large when the neighborhoods are not too large, because most of the random graph corresponds to between-neighborhood subgraphs and $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$ impose the same probability law on between-neighborhood subgraphs. To make the informal statements about the sizes of neighborhoods and the size of the deviation $\max_{\boldsymbol{z}} |\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X})|$ more precise, we compare the size of the deviation $\max_{\boldsymbol{z}} |\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X})|$ to the expected logliklihood function $\mathbb{E}\,\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star,\boldsymbol{z}^\star)}(\boldsymbol{X})$, which is a convenient measure of the size of the random graph. We note that, while it is tempting to believe that the size of the random graph is equal to the number of possible edges $\binom{n}{2}$, the size of the random graph depends on the sparsity of the random graph. The notion of sparsity of random graphs is motivated by the observation that most real-world networks are sparse in the sense that the observed number of edges is much smaller than the number of possible edges $\binom{n}{2}$. We call random graphs dense when $\mathbb{E}\,\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star,\boldsymbol{z}^\star)}(\boldsymbol{X})$ grows as $\binom{n}{2}$ and sparse otherwise. It is worth noting that the classic definition of sparsity is based on the expectation of the sufficient statistic of classic random graphs (Bollobás, 1998)—i.e., the expected number of edges—but in more general models it is desirable to base the definition of sparsity on all sufficient statistics and $\mathbb{E}\,\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star,\boldsymbol{z}^\star)}(\boldsymbol{X})$ is a convenient choice. As a result, the size $c\,(1+\delta)^{1/2}\,\|\mathcal{A}\|_\infty^2\, n_{\max}^2\, n\,(\log n)^{3/2}$ of the deviation $\max_{\boldsymbol{z}} |\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X})|$

is small relative to the size of the random graph $\mathbb{E}\ \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star,\boldsymbol{z}^\star)}(\boldsymbol{X})$ as long as $n_{\max}$ satisfies

$$n_{\max} \ \leq \ c_0 \ \left(\frac{\mathbb{E}\ \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star,\boldsymbol{z}^\star)}(\boldsymbol{X})}{\|\mathcal{A}\|_\infty^2 \ n \ (\log n)^{3/2}} \right)^{1/2}, \quad c_0 > 0.$$

If, e.g., the random graph is dense, then $n_{\max}$ must satisfy $n_{\max} \leq c_0 \ n^{1/2} \ / \ \|\mathcal{A}\|_\infty \ (\log n)^{3/4}$. If the random graph is sparse in the sense that $\mathbb{E}\ \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star,\boldsymbol{z}^\star)}(\boldsymbol{X})$ grows as $\|\mathcal{A}\|_\infty^2 \ n \ (\log n)^{3/2}$—i.e., $\mathbb{E}\ \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}^\star,\boldsymbol{z}^\star)}(\boldsymbol{X})$ grows faster than $n \log n$, which is the rate of growth of the expected loglikelihood function at the so-called threshold of connectivity of random graphs with independent and identically distributed edge variables (Bollobás, 1998)—then $n_{\max}$ must satisfy $n_{\max} \leq c_0$. These considerations suggest that as long as the neighborhoods are not too large and the random graph is not too sparse—i.e., the random graph is above the so-called threshold of connectivity—updates of $\boldsymbol{z}$ given $\boldsymbol{\theta}$ can be based on $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ rather than $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$.

3.4 Two-step likelihood-based approach

We propose a two-step likelihood-based approach that takes advantage of the theoretical results of Section 3.3 and enables massive-scale estimation of exponential-family random graph models with local dependence.

To describe the two-step likelihood-based approach, assume that $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ is the observed value of a random variable $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ with distribution

$$\boldsymbol{Z}_i \ \overset{\text{iid}}{\sim} \ \text{Multinomial}(1, \ \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)), \quad i = 1, \ldots, n.$$

It is natural to base statistical inference on the observed-data likelihood function

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\pi}) \ = \ \sum_{\boldsymbol{z} \in \mathbb{Z}} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) \, p_{\boldsymbol{\pi}}(\boldsymbol{z}).$$

The problem is that $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\pi})$ is intractable, because $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ is intractable and the set $\mathbb{Z}$ contains $\exp(n \log K)$ elements.

The first problem can be solved by taking advantage of the theoretical results of Section 3.3, which suggest that $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ can be approximated by $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ provided that the neighborhoods are not too large and the random graph is not too sparse. A complication is that $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$

and $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$ may not be close when the neighborhoods are large, i.e., when $\boldsymbol{z} \in \mathbb{Z} \setminus \mathbb{S}$. However, the basic inequality

$$\sum_{\boldsymbol{z}\in\mathbb{S}} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) \;\leq\; \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\pi}) \;\leq\; \sum_{\boldsymbol{z}\in\mathbb{S}} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) + \mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} \in \mathbb{Z} \setminus \mathbb{S})$$

suggests that as long as the event $\boldsymbol{Z} \in \mathbb{Z} \setminus \mathbb{S}$ is a rare event in the sense that $\mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} \in \mathbb{Z} \setminus \mathbb{S})$ is close to 0, $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\pi})$ can be approximated by $\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2{=}\mathbf{0},\boldsymbol{\pi})$:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\pi}) \;&=\; \sum_{\boldsymbol{z}\in\mathbb{Z}} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) \;\approx\; \sum_{\boldsymbol{z}\in\mathbb{S}} p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) \\
&\approx\; \sum_{\boldsymbol{z}\in\mathbb{S}} p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) \;\approx\; \sum_{\boldsymbol{z}\in\mathbb{Z}} p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})\, p_{\boldsymbol{\pi}}(\boldsymbol{z}) \;=\; \mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2{=}\mathbf{0},\boldsymbol{\pi}).
\end{aligned}$$

We note that the assumption that $\boldsymbol{Z} \in \mathbb{Z} \setminus \mathbb{S}$ is a rare event makes sense in a wide range of applications, because communities in real-world networks tend to be small (see, e.g., the discussion of Rohe et al., 2011). Therefore, as long as $\boldsymbol{Z} \in \mathbb{Z} \setminus \mathbb{S}$ is a rare event, we can base statistical inference concerning the neighborhood structure on $\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = \mathbf{0}, \boldsymbol{\pi})$ rather than $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\pi})$. To simplify the notation, we write henceforth $\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\pi})$ instead of $\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2{=}\mathbf{0},\boldsymbol{\pi})$.

The second problem can be solved by methods developed for stochastic block models, because $\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\pi})$ is the observed-data likelihood function of a stochastic block model. There are many methods that could be used, such as profile likelihood (Bickel and Chen, 2009), pseudo-likelihood (Amini et al., 2013), spectral clustering (Rohe et al., 2011), and variational methods (Daudin et al., 2008; Vu et al., 2013). Among these methods, we found that the variational methods of Vu et al. (2013) work best in practice. In addition, the variational methods of Vu et al. (2013) have the advantage of being able to estimate stochastic block models from networks with hundreds of thousands of nodes due to a running time of $O(n)$ for sparse random graphs and $O(n^2)$ for dense random graphs (Vu et al., 2013). Some consistency and asymptotic normality results for variational methods for stochastic block models were established by Celisse et al. (2012) and Bickel et al. (2013).

Variational methods approximate $\ell(\boldsymbol{\theta}_1,\boldsymbol{\pi}) = \log \mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\pi})$ by introducing an auxiliary distribu-

tion $a(\boldsymbol{z})$ with support $\mathbb{Z}$ and lower bound $\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi})$ by using Jensen's ineqality:

$$\begin{aligned}\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi}) &= \log \sum_{\boldsymbol{z} \in \mathbb{Z}} a(\boldsymbol{z}) \, \frac{p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = \mathbf{0}, \boldsymbol{z})}(\boldsymbol{x}) \, p_{\boldsymbol{\pi}}(\boldsymbol{z})}{a(\boldsymbol{z})} \\ &\geq \sum_{\boldsymbol{z} \in \mathbb{Z}} a(\boldsymbol{z}) \, \log \frac{p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = \mathbf{0}, \boldsymbol{z})}(\boldsymbol{x}) \, p_{\boldsymbol{\pi}}(\boldsymbol{z})}{a(\boldsymbol{z})} \quad \stackrel{\text{def}}{=} \quad \hat{\ell}(\boldsymbol{\theta}_1, \boldsymbol{\pi}).\end{aligned}$$

Each auxiliary distribution with support $\mathbb{Z}$ gives rise to a lower bound on $\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi})$. To choose the best auxiliary distribution—i.e., the auxiliary distribution that gives rise to the tightest lower bound on $\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi})$—we choose a family of auxiliary distributions and select the best member of the family. In practice, an important consideration is that the resulting lower bound is tractable. Therefore, we confine attention to a family of auxiliary distributions under which the resulting lower bounds are tractable. A natural choice is given by a family of auxiliary distributions under which the neighborhood memberships are independent:

$$\boldsymbol{Z}_i \quad \stackrel{\text{ind}}{\sim} \quad \text{Multinomial}(1, \, \boldsymbol{\alpha}_i = (\alpha_{i,1}, \ldots, \alpha_{i,K})), \quad i = 1, \ldots, n.$$

By the independence of neighborhood memberships under the auxiliary distribution, one obtains the following tractable lower bound on $\ell(\boldsymbol{\theta}_1, \boldsymbol{\pi})$ (Vu et al., 2013):

$$\begin{aligned}&\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}) \stackrel{\text{def}}{=} \sum_{\boldsymbol{z} \in \mathbb{Z}} a_{\boldsymbol{\alpha}}(\boldsymbol{z}) \, \log \frac{p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = \mathbf{0}, \boldsymbol{z})}(\boldsymbol{x}) \, p_{\boldsymbol{\pi}}(\boldsymbol{z})}{a_{\boldsymbol{\alpha}}(\boldsymbol{z})} \\ = \quad &\sum_{i<j}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_{i,k} \, \alpha_{j,l} \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = \mathbf{0}, z_{i,k}=1, \, z_{j,l}=1, \boldsymbol{z}_{-i,j})}(x_{i,j}) + \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha_{i,k} \, (\log \pi_k - \log \alpha_{i,k}),\end{aligned}$$

where $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = \mathbf{0}, z_{i,k}=1, \, z_{j,l}=1, \boldsymbol{z}_{-i,j})}(x_{i,j})$ denotes the marginal probability mass function of $X_{i,j}$ and $\boldsymbol{z}_{-i,j}$ the neighborhood memberships of all nodes excluding nodes i and j.

In practice, we obtain the best lower bound by maximizing $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ with respect to $\boldsymbol{\alpha}$. Direct maximization of $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ with respect to $\boldsymbol{\alpha}$ is possible but inconvenient, because $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ contains products of $\alpha_{i,k}$ and $\alpha_{j,l}$. As a consequence, a fixed-point update of $\alpha_{i,k}$ would depend on $(n-1) \, K$ other terms $\alpha_{j,l}$ and hence fixed-point updates tend to be time-consuming and get stuck in local maxima (Vu et al., 2013). An elegant approach to alleviate the problem is to use minorization-maximization methods (Hunter and Lange, 2004). Such methods construct a minorizing function that approximates $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ but is easier to maximize than $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$. A

function $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ of $\boldsymbol{\alpha}$ minorizes $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ at point $\boldsymbol{\alpha}^{(t)}$ at iteration t of an iterative algorithm for maximizing $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ if

$$\begin{aligned} M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)}) &\leq \hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}) \qquad \text{for all} \quad \boldsymbol{\alpha}, \\ M(\boldsymbol{\alpha}^{(t)}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)}) &= \hat{\ell}(\boldsymbol{\alpha}^{(t)}; \boldsymbol{\theta}_1, \boldsymbol{\pi}), \end{aligned}$$

where $\boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)}$ are fixed. In other words, $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ is bounded above by $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ for all $\boldsymbol{\alpha}$ and touches $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(t)}$. As a result, increasing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ with respect to $\boldsymbol{\alpha}$ increases $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$. Vu et al. (2013) showed that the following function minorizes $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ at point $\boldsymbol{\alpha}^{(t)}$:

$$\begin{aligned} M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)}) &= \sum_{i<j}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \left(\alpha_{i,k}^2 \, \frac{\alpha_{j,l}^{(t)}}{2\,\alpha_{i,k}^{(t)}} + \alpha_{j,l}^2 \, \frac{\alpha_{i,k}^{(t)}}{2\,\alpha_{j,l}^{(t)}} \right) \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = \mathbf{0}, z_{i,k}=1,\, z_{j,l}=1, \boldsymbol{z}_{-i,j})}(x_{i,j}) \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha_{i,k} \left[\log \pi_k^{(t)} - \log \alpha_{i,k}^{(t)} + \left(1 - \frac{\alpha_{i,k}}{\alpha_{i,k}^{(t)}} \right) \right]. \end{aligned}$$

The minorizing function $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ is easier to maximize than $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$, because it replaces the products of $\alpha_{i,k}$ and $\alpha_{j,l}$ by sums of $\alpha_{i,k}^2$ and $\alpha_{j,l}^2$. An additional advantage is that the maximization of $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ amounts to n quadratic programming problems, which can be solved in parallel.

We therefore propose a two-step likelihood-based approach as described in Table 3.1. We discuss the two steps below and conclude with some comments on parallel computing.

Step 1 The first step estimates $\boldsymbol{z}$ based on $\boldsymbol{\alpha}$. We do so by increasing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ with respect to $\boldsymbol{\alpha}_i$ subject to the constraints $\alpha_{i,k} \geq 0$ and $\sum_{k=1}^{K} \alpha_{i,k} = 1$ $(i = 1, \ldots, n)$. We increase rather than maximize $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$, because maximizing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ is more time-consuming and algorithms maximizing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ are more prone to end up in local maxima than algorithms increasing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$. Since $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ and $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi}, \boldsymbol{\alpha}^{(t)})$ depend on $\boldsymbol{\theta}_1$ and $\boldsymbol{\pi}$ and both are unknown, we iterate between updates of $\boldsymbol{\alpha}$ and updates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\pi}$. The updates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\pi}$ are based on maximizing $\hat{\ell}(\boldsymbol{\alpha}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ with respect to $\boldsymbol{\theta}_1$ and $\boldsymbol{\pi}$ and are identical to the updates of Vu et al. (2013), because $\boldsymbol{\theta}_2 = \mathbf{0}$ reduces the model to a stochastic block model. As a

1. Estimate $\boldsymbol{z}$ along with $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_1$ by iterating:

   1.1 Update $\boldsymbol{\alpha}$ by increasing $M(\boldsymbol{\alpha}; \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\alpha}^{(t)})$ with respect to $\boldsymbol{\alpha}_i$ subject to $\alpha_{i,k} \geq 0$ and $\sum_{k=1}^{K} \alpha_{i,k} = 1$ and denote the update by $\boldsymbol{\alpha}_i^{(t+1)}$ $(i = 1, \ldots, n)$.

   1.2 Update $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_1$ by maximizing $\hat{\ell}(\boldsymbol{\alpha}^{(t+1)}; \boldsymbol{\theta}_1, \boldsymbol{\pi})$ with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_1$:

   — Update $\pi_k^{(t+1)} = (1/n) \sum_{i=1}^{n} \alpha_{i,k}^{(t+1)}$, $k = 1, \ldots, K$.

   — Update $\boldsymbol{\theta}_1^{(t+1)} = \arg\max_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} \hat{\ell}(\boldsymbol{\alpha}^{(t+1)}; \boldsymbol{\theta}_1, \boldsymbol{\pi}^{(t+1)})$.

   Upon convergence, we estimate the neighborhood memberhip indicators by $\widehat{z}_{i,k} = 1$ if $k = \arg\max_{1 \leq l \leq K} \widehat{\alpha}_{i,l}$ and $\widehat{z}_{i,k} = 0$ otherwise $(i = 1, \ldots, n, k = 1, \ldots, K)$, where $\widehat{\boldsymbol{\alpha}}$ denotes the final value of $\boldsymbol{\alpha}$.

2. Estimate $\boldsymbol{\theta}$ given $\widehat{\boldsymbol{z}}$ by $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{\ell}_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta})$.

Table 3.1 : Two-step likelihood-based approach.

convergence criterion, we use

$$\frac{|\hat{\ell}(\boldsymbol{\alpha}^{(t+1)}; \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\pi}^{(t+1)}) - \hat{\ell}(\boldsymbol{\alpha}^{(t)}; \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\pi}^{(t)})|}{\hat{\ell}(\boldsymbol{\alpha}^{(t+1)}; \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\pi}^{(t+1)})} \quad < \quad \gamma,$$

where $\gamma > 0$ is a small constant. Upon convergence, we estimate the neighborhood memberhip indicators by $\widehat{z}_{i,k} = 1$ if $k = \arg\max_{1 \leq l \leq K} \widehat{\alpha}_{i,l}$ and $\widehat{z}_{i,k} = 0$ otherwise $(i = 1, \ldots, n,\ k = 1, \ldots, K)$, where $\widehat{\boldsymbol{\alpha}}$ denotes the final value of $\boldsymbol{\alpha}$.

Step 2 We estimate $\boldsymbol{\theta}$ given $\widehat{\boldsymbol{z}}$ by using the Monte Carlo maximum likelihood methods (Hunter and Handcock, 2006). Monte Carlo maximum likelihood methods exploit the fact that the loglikelihood function induced by $\widehat{\boldsymbol{z}}$, which is defined by

$$\ell_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta}) \quad = \quad \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \widehat{\boldsymbol{z}})}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}})}(\boldsymbol{x}),$$

can be written as

$$\ell_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta}) \;=\; \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \widehat{\boldsymbol{z}}) - \boldsymbol{\eta}(\boldsymbol{\theta}_0,\, \widehat{\boldsymbol{z}}),\, s(\boldsymbol{x})\rangle - \log \mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}_0,\, \widehat{\boldsymbol{z}})} \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \widehat{\boldsymbol{z}}) - \boldsymbol{\eta}(\boldsymbol{\theta}_0,\, \widehat{\boldsymbol{z}}),\, s(\boldsymbol{X})\rangle),$$

where $\boldsymbol{\theta}_0$ is a fixed parameter vector (e.g., $\boldsymbol{\theta}_0$ may be an educated guess of $\boldsymbol{\theta}^\star$). In general, the expectation $\mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}_0,\, \widehat{\boldsymbol{z}})}$ is intractable, but it can be estimated by a Monte Carlo sample average based on a Monte Carlo sample of graphs generated under $\boldsymbol{\eta}(\boldsymbol{\theta}_0,\, \widehat{\boldsymbol{z}})$. Therefore, we can approximate $\ell_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta})$ by

$$\hat{\ell}_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta}) \;=\; \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \widehat{\boldsymbol{z}}) - \boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}}), s(\boldsymbol{x})\rangle - \log \widehat{\mathbb{E}}_{\boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}})} \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \widehat{\boldsymbol{z}}) - \boldsymbol{\eta}(\boldsymbol{\theta}_0, \widehat{\boldsymbol{z}}), s(\boldsymbol{X})\rangle),$$

where $\widehat{\mathbb{E}}_{\boldsymbol{\eta}(\boldsymbol{\theta}_0,\, \widehat{\boldsymbol{z}})}$ is a Monte Carlo approximation of $\mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}_0,\, \widehat{\boldsymbol{z}})}$ based on a Monte Carlo sample of graphs generated by using $\boldsymbol{\eta}(\boldsymbol{\theta}_0,\, \widehat{\boldsymbol{z}})$. Hence $\boldsymbol{\theta}$ given $\widehat{\boldsymbol{z}}$ can be estimated by

$$\widehat{\boldsymbol{\theta}} \;=\; \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{\ell}_{\widehat{\boldsymbol{z}}}(\boldsymbol{\theta}).$$

Additional details on Monte Carlo maximum likelihood methods can be found in Hunter and Handcock (2006). We note that the local dependence of the model facilitates parallel computing, which is discussed in the following paragraph. Standard errors of $\widehat{\boldsymbol{\theta}}$ can be based on the estimated Fisher information matrix, although such standard errors are conditional on the estimated neighborhood structure $\widehat{\boldsymbol{z}}$ and therefore do not reflect the uncertainty about $\widehat{\boldsymbol{z}}$. A parametric bootstrap approach would be an interesting approach for capturing the additional uncertainty due to $\widehat{\boldsymbol{z}}$, but it would be time-consuming.

Parallel computing In Step 1, the maximization of the minorizing function amounts to n quadratic programming problems, which can be solved in parallel. In Step 2, the local dependence induced by the model implies that the contributions of the between- and within-neighborhood subgraphs to the loglikelihood function and its gradient and Hessian can be computed in parallel. Hence both steps can be implemented in parallel, which suggests that the two-step likelihood-based method can be used on a massive scale as long as the neighborhoods are not too large and multi-core computers or computing clusters are available.

	Two-step likelihood-based approach	Bayesian approach
$n = 30$, $K = 3$, balanced	46.6	14,735.1
$n = 30$, $K = 3$, unbalanced	48.3	17,853.2

Table 3.2 : Computing time in seconds: two-step likelihood-based approach versus Bayesian approach. The two-step likelihood-based approach did not exploit parallel computing in Step 1, but exploited 3 cores in Step 2 to deal with the $K = 3$ within-neighborhood subgraphs.

3.5 Simulation results

We first compare the two-step likelihood-based approach to the Bayesian approach of Schweinberger and Handcock (2015), which is the gold standard for small networks, and then assess the performance of the two-step likelihood-based approach on large networks. Throughout, we focus on undirected random graphs with sample space $\mathbb{X} = \{0, 1\}^{\binom{n}{2}}$.

To compare the two-step likelihood-based approach to the Bayesian approach, we focus on small random graphs with $n = 30$ nodes and $K = 3$ neighborhoods, because the Bayesian approach is too time-consuming to be applied to large networks. We consider two cases. In the first case, called the balanced case, all 3 neighborhoods contain 10 nodes. In the second case, called the unbalanced case, the 3 neighborhoods contain 5, 10, and 15 nodes, respectively. In addition, we compare the two-step likelihood-based approach to the spectral clustering method of Lei and Rinaldo (2015), which ignores the model structure and estimates the neighborhood structure by spectral clustering; note that spectral clustering is an alternative to the variational methods in the first step of the two-step likelihood-based approach, as mentioned in Section 3.4. To assess the performance of the two-step likelihood-based approach on large networks, we focus on random graphs with $n = 2{,}500$ nodes in $K = 100$ neighborhoods. Once again, we consider two cases, the balanced case with 100 neighborhoods of size 25 and the unbalanced case with 20 neighborhoods of sizes 15, 20, 25, 30, and 35, respectively.

In each scenario, we generate 500 graphs from the exponential-family random graph model

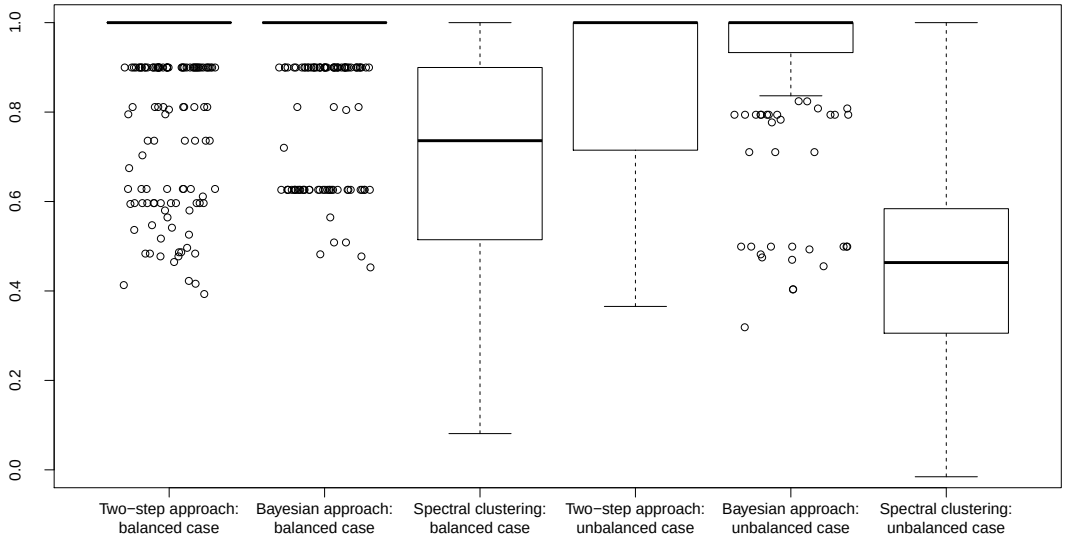


Figure 3.1 : Agreement of estimated and data-generating neighborhood structure in terms of Yule's ϕ-coefficient (value of 1 indicates perfect agreement) based on 500 simulated graphs with $n = 30$ nodes and $K = 3$ neighborhoods in the balanced and unbalanced case.

with within-neighborhood edges

$$s_{1,k,k}(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,k}$$

and transitive edges

$$s_{2,k,k}(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,k}\, \max_{h \neq i,j} x_{i,h}\, x_{j,h}\, z_{h,k}$$

and between-neighborhood edges

$$s_{k,l}(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i<j}^{n} x_{i,j}\, z_{i,k}\, z_{j,l}$$

as sufficient statistics and natural parameters $\eta_{1,k,k}(\boldsymbol{\theta}, \boldsymbol{z}) = \theta_1 \log n_k(\boldsymbol{z})$, $\eta_{2,k,k}(\boldsymbol{\theta}, \boldsymbol{z}) = \theta_2 \log n_k(\boldsymbol{z})$, and $\eta_{1,k,l}(\boldsymbol{\theta}, \boldsymbol{z}) = \theta_3 \log n$, where $n_k(\boldsymbol{z})$ is the size of neighborhood k under $\boldsymbol{z} \in \mathbb{Z}$. We use size-dependent parameterizations, because we do not want to force small and large neighborhoods to have the same natural parameters. The choice of the size-dependent parameterization used above is motivated by the sparsity of random graphs: e.g., in the case of classic random graphs which

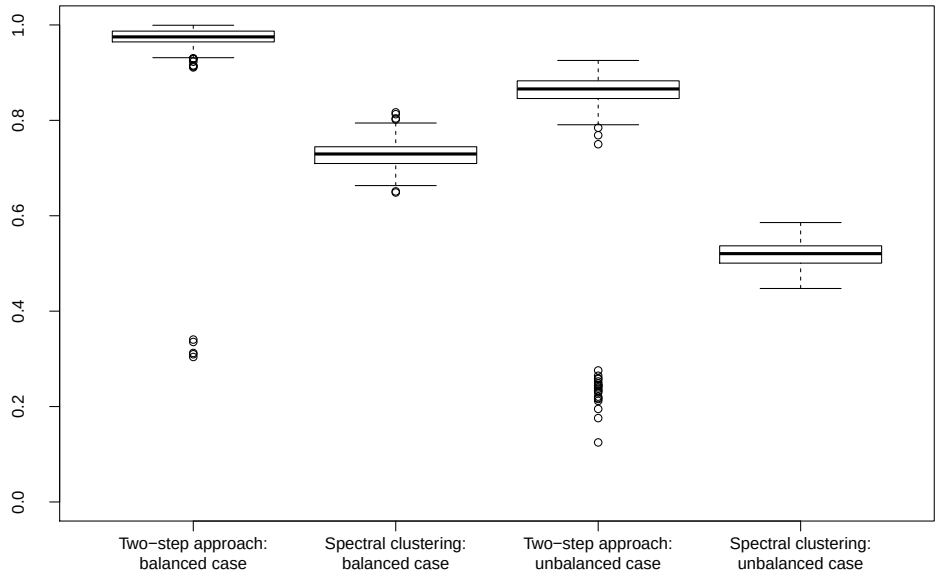


Figure 3.2 : Agreement of estimated and data-generating neighborhood structure in terms of Yule's ϕ-coefficient (value of 1 indicates perfect agreement) based on 500 simulated graphs with $n =$ 2,500 nodes and $K = 100$ neighborhoods in the balanced and unbalanced case.

assume that edges are independent Bernoulli(μ) random variables, it makes sense to assume that there exist $c > 0$ and $0 \leq \alpha \leq 1$ such that the expected number of edges of each node—which is given by $(n-1)\,\mu$—is bounded above by $c\ n^{\alpha}$, because real-world networks are sparse. As a consequence, μ should be of order n^{θ} and $\eta = \text{logit}(\mu)$ should be of order $\log n^{\theta} = \theta \log n$, where $\theta = \alpha - 1 < 0$. In more general exponential-family random graph models with edge terms as well as other model terms, all model terms should scale as the edge term, so that no model term can dominate any other model term. These considerations suggest that the natural parameters of within-neighborhood subgraphs should be of the form $\eta_{i,k,k}(\boldsymbol{\theta}, \boldsymbol{z}) = \theta_i \log n_k(\boldsymbol{z})$ ($i = 1, 2$, $k = 1, \ldots, K$) and the natural parameters of between-neighborhood subgraphs should be of the form $\eta_{1,k,l}(\boldsymbol{\theta}, \boldsymbol{z}) = \theta_3 \log n$ ($k < l = 1, \ldots, K$). We note that the size-dependent parameterization imposes a form of local sparsity on within-neighborhood subgraphs and a form of global sparsity on between-neighborhood subgraphs. The strength of sparsity depends on the size of the graph as

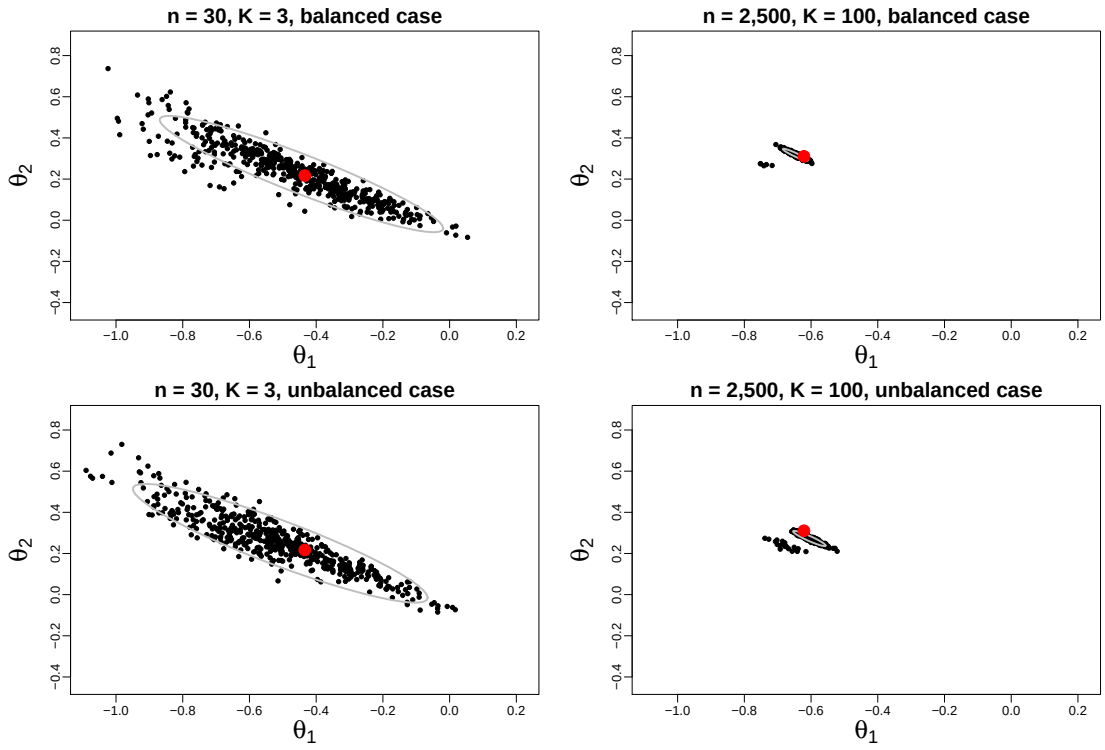


Figure 3.3 : Estimates of parameter vector $\boldsymbol{\theta}$ based on small and large networks in the balanced and unbalanced case; note that $\boldsymbol{\theta}$ should not be confused with the size-dependent natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{z})$. The red circles indicate the data-generating parameter vectors. The ellipses correspond to 95% quantiles of the fitted bivariate t-distribution.

well as parameter vector $\boldsymbol{\theta}$.

We compare the three methods described above in terms of neighborhood recovery by using Yule's ϕ-coefficient:

$$\phi(\boldsymbol{z}^\star, \boldsymbol{z}) \quad = \quad \frac{n_{0,0}\, n_{1,1} - n_{0,1}\, n_{1,0}}{\sqrt{(n_{0,0} + n_{0,1})\,(n_{1,0} + n_{1,1})\,(n_{0,0} + n_{1,0})\,(n_{0,1} + n_{1,1})}}, \tag{3.3}$$

where

$$n_{a,b} \quad = \quad n_{a,b}(\boldsymbol{z}^\star, \boldsymbol{z}) \quad = \quad \sum_{i<j}^{n} \mathbb{1}(\mathbb{1}(\boldsymbol{z}_i^\star = \boldsymbol{z}_j^\star) = a)\ \mathbb{1}(\mathbb{1}(\boldsymbol{z}_i = \boldsymbol{z}_j) = b), \quad a, b \in \{0, 1\}.$$

Here, $\mathbb{1}(.)$ is an indicator function, which is 1 if the statement in parentheses is true and is 0 otherwise. It is worth noting that Yule's ϕ-coefficient is invariant to the labeling of the neighborhoods and is bounded above by 1, where 1 indicates perfect agreement of the data-generating and estimated neighborhood structure.

In the small-network scenario, we generate data by using between-neighborhood natural parameters $\eta_{1,k,l}(\boldsymbol{\theta}^\star, \boldsymbol{z}^\star) = -.882 \log n$ and within-neighborhood natural parameters $\eta_{1,k,k}(\boldsymbol{\theta}^\star, \boldsymbol{z}^\star) = -.434 \log n_k(\boldsymbol{z}^\star)$ and $\eta_{2,k,k}(\boldsymbol{\theta}^\star, \boldsymbol{z}^\star) = .217 \log n_k(\boldsymbol{z}^\star)$. According to Figure 3.1, the two-step likelihood-based approach is almost as good as the Bayesian approach in terms of neighborhood recovery in the balanced case but worse in the unbalanced case. The worse performance in the unbalanced case may be due to the fact that there are smaller neighborhoods in the unbalanced case than in the balanced case and recovering small neighborhoods is more challenging than recovering large neighborhoods. However, while the Bayesian approach has a small advantage in the unbalanced case, Table 3.2 shows that the cost of the small improvement in neighborhood recovery is excessive: the computing time of the Bayesian approach is 370 times higher than the computing time of the two-step likelihood-based approach.

In the second scenario, we generate data by using $\eta_{1,k,k}(\boldsymbol{\theta}^\star, \boldsymbol{z}^\star) = -.621 \log n_k(\boldsymbol{z}^\star)$, $\eta_{2,k,k}(\boldsymbol{\theta}^\star, \boldsymbol{z}^\star) = .311 \log n_k(\boldsymbol{z}^\star)$, and $\eta_{1,k,l}(\boldsymbol{\theta}^\star, \boldsymbol{z}^\star) = -.511 \log n$. Figure 3.2 shows that the two-step likelihood-based approach outperforms spectral clustering in terms of neighborhood recovery in most cases. In the few cases where spectral clustering outperforms the two-step likelihood-based approach, the variational algorithms may have been trapped in local maxima.

Last, but not least, we assess the performance of the two-step likelihood-based approach in terms of parameter recovery. Figure 3.3 shows that the estimated parameters are close to the data-generating parameter vectors, and more so when the number of neighborhoods is large. Once again, in the few cases where estimates are far from the data-generating parameter vector, the variational algorithms may have been trapped in local maxima. In such cases, the neighborhood recovery can be poor, which in turn affects the parameter recovery.

3.6 Application to large Amazon product network

We use the two-step likelihood-based approach to shed light on the complex structure of a large Amazon product network. The data on the Amazon product network were collected by Yang and Leskovec (2015) and can be downloaded from the website

`http://snap.stanford.edu/data/com-Amazon.html`

The network consists of products listed at `www.amazon.com`. Two products i and j are connected by an edge if i and j are frequently purchased together according to the "Customers Who Bought This Item Also Bought" feature at `www.amazon.com`. Amazon assigns all products to categories, which we consider to be ground-truth neighborhoods. We use a subset of the network consisting of the top 500 non-overlapping categories with 10 to 80 products, where the ranking of categories is based on Yang and Leskovec (2015). The resulting network consists of 10,448 products and 33,537 edges and can be found in the supplementary archive.

To model the Amazon product network, we take advantage of curved exponential-family random graph models. To capture the complex structure of within-neighborhood subgraphs, we use within-neighborhood edge terms, geometrically weighted degree terms, and geometrically weighted edgewise shared partner terms. The natural parameters of the within-neighborhood edge terms are given by

$$\eta_{1,k,k}(\boldsymbol{\theta}, \boldsymbol{z}) \;\; = \;\; \theta_1 \, \log n_k(\boldsymbol{z}).$$

The within-neighborhood geometrically weighted degree terms are based on the number of products with t edges in neighborhood $\mathcal{A}_k$. The natural parameters of within-neighborhood geometrically weighted degree terms are given by

$$\eta_{2,k,k,t}(\boldsymbol{\theta}, \boldsymbol{z}) \;\; = \;\; \theta_2 \, \log n_k(\boldsymbol{z}) \, \exp(\theta_3) \, \left[1 - (1 - \exp(-\theta_3))^t\right], \;\; t = 1, \ldots, n_k(\boldsymbol{z}) - 1.$$

The within-neighborhood geometrically weighted edgewise shared partner terms are based on the number of connected pairs of products i and j in neighborhood $\mathcal{A}_k$ such that i and j have t shared partners in neighborhood $\mathcal{A}_k$. The natural parameters of the within-neighborhood geometrically weighted edgewise shared partner terms are given by

$$\eta_{3,k,k,t}(\boldsymbol{\theta}, \boldsymbol{z}) \;\; = \;\; \theta_4 \, \log n_k(\boldsymbol{z}) \, \exp(\theta_5) \, \left[1 - (1 - \exp(-\theta_5))^t\right], \;\; t = 1, \ldots, n_k(\boldsymbol{z}) - 2.$$

To reduce computing time, it is convenient to truncate the two geometrically weighted model terms by setting $\eta_{2,k,k,t}(\boldsymbol{\theta}, \boldsymbol{z}) = 0$, $t = 21, \ldots, n_k(\boldsymbol{z}) - 1$, and $\eta_{3,k,k,t}(\boldsymbol{\theta}, \boldsymbol{z}) = 0$, $t = 13, \ldots, n_k(\boldsymbol{z}) - 2$.

The two thresholds 21 and 13 are motivated by the fact that no product has 21 or more edges and less than 1% of all pairs of products has 13 or more edgewise shared partners. Last, but not least, the natural parameters of the between-neighborhood edge terms are given by

$$\eta_{1,k,l}(\boldsymbol{\theta}, \boldsymbol{z}) \;\; = \;\; \theta_6 \, \log n, \quad k < l.$$

The resulting exponential family is a curved exponential family (Hunter and Handcock, 2006), because the natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{z})$ of the exponential family is a nonlinear function of $\boldsymbol{\theta}$ given $\boldsymbol{z} \in \mathbb{Z}$. In addition, the natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{z})$ is size-dependent, because we do not want to force small and large neighborhoods to have the same natural parameters, as explained in Section 3.5. It is worth noting that the inclusion of the geometrically weighted degree terms helps model the connectivity of the network, while the inclusion of the geometrically weighted edgewise shared partner terms helps capture transitivity, i.e., the tendency of products i and k to be co-purchased when products i and j and products j and k tend to be co-purchased. Transitivity can arise when, e.g., (a) three products are similar (e.g., three books on the same topic); (b) three products are dissimilar but complement each other (e.g., a bicycle helmet, head light, and tail light); (c) three products, either similar or dissimilar, were produced by the same source (e.g., three books written by the same author); and (d) when customers become aware that products i and j and products j and k tend to be co-purchased, some customers might start co-purchasing i and k even though Amazon might not recommend co-purchases of i and k: e.g., when a new product i is introduced (e.g., a novel) and product i is known to be related to product j (e.g., a novel by the same author), and product j tends to be co-purchased with product k (e.g., a classic novel), then customers might start co-purchasing i and k even though Amazon might not recommend co-purchases of i and k.

Since we know the number of ground-truth neighborhoods, we set $K = 500$ and estimate the neighborhood structure by using the two-step likelihood-based approach. To assess the performance of the two-step likelihood-based approach in terms of neighborhood recovery, we use Yule's ϕ-coefficient. Yule's ϕ-coefficient turns out to be $.964$, which indicates near-perfect recovery of the ground-truth neighborhood structure. The Monte Carlo maximum likelihood estimates

Term	Estimate	S.E.	Estimate	S.E.
Within-neighborhood edges θ_1	$-.368$	.002	-1.403	.014
Within-neighborhood degrees θ_2			1.086	.020
Within-neighborhood degrees θ_3			.760	.027
Within-neighborhood shared partners θ_4			.291	.003
Within-neighborhood shared partners θ_5			1.161	.004
Between-neighborhood edges θ_6	-1.197	$< .001$	-1.197	$< .001$

Table 3.3 : Monte Carlo maximum likelihood estimates and standard errors (S.E.) of $\theta_1, \ldots, \theta_6$ estimated from the Amazon product network with 10,448 products; note that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)$ should not be confused with the size-dependent natural parameter vector $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{z})$.

and standard errors of $\theta_1, \ldots, \theta_6$ are shown in Table 3.3 and suggest that there is evidence for transitivity. The observed tendency toward transitivity has at least two advantages in practice. First, it suggests that Amazon might be able to improve recommendations by recommending customers of product i to purchase product k provided that i and k are connected to at least one other product, even though products i and k might not have been co-purchased in the past (see, e.g., example (d) above: i or k or both might be new products known to be related to existing products). Second, it suggests that Amazon might be able to partition large categories into small subcategories based on the transitive structure within categories.

To demonstrate that the curved exponential-family random graph model considered here can capture structural features of networks that simple models, such as stochastic block models, cannot capture, we compare the goodness-of-fit of the curved exponential-family random graph model to the goodness-of-fit of stochastic block models. Since the two models impose the same probability law on between-neighborhood subgraphs, it is natural to compare the two models in terms of goodness-of-fit with respect to within-neighborhood subgraphs. We assess the goodness-of-fit of the two models in terms of the within-neighborhood geodesic distances of pairs of products, i.e., the length of the shortest path between pairs of products in the same neighborhood; the numbers of

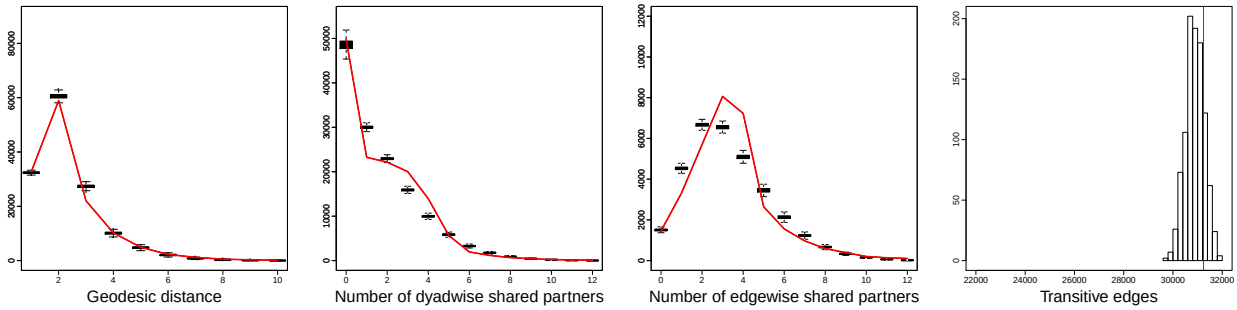


Figure 3.4 : Amazon product network with 10,448 products: goodness-of-fit of curved exponential-family random graph model. The red lines indicate observed values of statistics.

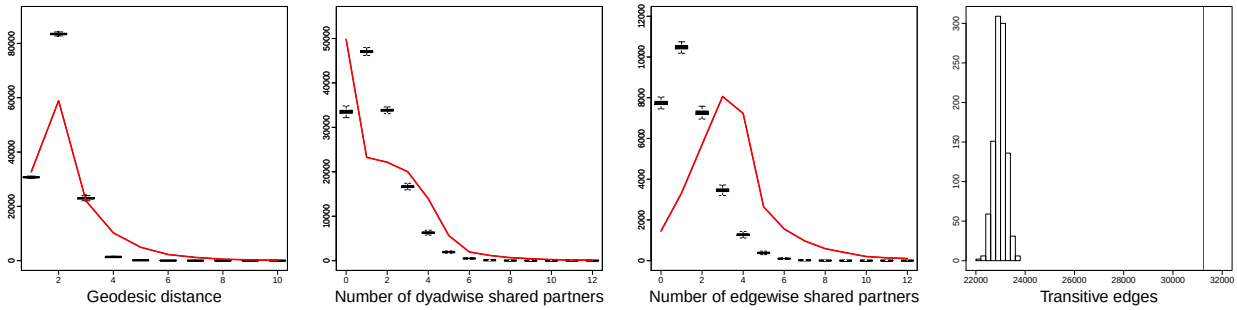


Figure 3.5 : Amazon product network with 10,448 products: goodness-of-fit of stochastic block models. The red lines indicate observed values of statistics.

within-neighborhood dyadwise shared partners, i.e., the number of unconnected or connected pairs of products with i shared partners in the same neighborhood; the numbers of within-neighborhood edgewise shared partners, i.e., the number of connected pairs of products with i shared partners in the same neighborhood; and the number of transitive edges, i.e., the number of pairs of products with at least one shared partner in the same neighborhood. Figures 3.4 and 3.5 compare the goodness-of-fit of the two models based on 1,000 graphs simulated from the estimated models. The figures suggest that the curved exponential-family random graph model considered here is superior to the stochastic block model in terms of both connectivity and transitivity.

3.7 Appendix: Proofs of Chapter 3

To prove Theorem 3.1, we need three additional results, Lemma 3.1 and Propositions 3.1 and 3.2. To state them, let

$$g(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{z}) \;\; = \;\; \log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x}) - \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x}),$$

where $g(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{z})$ is considered as a function of $\boldsymbol{x} \in \mathbb{X}$ for fixed $(\boldsymbol{\theta},\boldsymbol{z}) \in \boldsymbol{\Theta} \times \mathbb{Z}$. Observe that the expectation $\mathbb{E}\, s(\boldsymbol{X})$ exists (Brown, 1986, Theorem 2.2, pp. 34–35), because $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \boldsymbol{\Xi}$ and $\boldsymbol{\Xi} \subseteq \text{int}(\mathbb{N})$ is a subset of the interior $\text{int}(\mathbb{N})$ of the natural parameter space $\mathbb{N}$. Therefore, the expectations $\mathbb{E} \log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X})$ and $\mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})$ exist, because

$$\mathbb{E} \log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) \;\; = \;\; \langle \boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}),\, \mathbb{E}\, s(\boldsymbol{X}) \rangle - \psi(\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}))$$

and

$$\mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z}) \;\; = \;\; \mathbb{E} \log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \mathbb{E} \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{X}).$$

We first state Lemma 3.1 and Propositions 3.1 and 3.2 and then prove Theorem 3.1.

Lemma 3.1 *Suppose that a random graph $\boldsymbol{X}$ is governed by an exponential family with countable support $\mathbb{X}$ and local dependence. Let $f : \mathbb{X} \times \mathbb{Z} \mapsto \mathbb{R}$ be a function of within-neighborhood edge variables $(X_{i,j})^n_{i<j:\, \boldsymbol{z}_i=\boldsymbol{z}_j}$ that is Lipschitz with respect to the Hamming metric $d : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}^+_0$ with Lipschitz coefficient $\|f\|_{Lip} > 0$ and $\mathbb{E}\, f(\boldsymbol{X};\boldsymbol{z}) < \infty$. Then there exists $c > 0$ such that, for all $\boldsymbol{z} \in \mathbb{Z}$, all $n > 0$, and all $t > 0$,*

$$\mathbb{P}\left(|f(\boldsymbol{X};\boldsymbol{z}) - \mathbb{E}\, f(\boldsymbol{X};\boldsymbol{z})| \;\geq\; t\right) \quad \leq \quad 2\, \exp\left(-\frac{t^2}{c\, K\, n_{\max}(\boldsymbol{z})^2\, \|\mathcal{A}\|^4_\infty\, \|f\|^2_{Lip}}\right).$$

Proof. The proof of Lemma 3.1 follows the proof of Proposition 1 of Schweinberger and Stewart (2016) and is therefore omitted.

Proposition 3.1 *Suppose that a random graph is governed by an exponential-family random graph model with countable support $\mathbb{X}$ and local dependence satisfying conditions [C.1] and [C.2]. Let*

$\mathbb{S} \subseteq \mathbb{Z}$ be a subset of neighborhood structures such that $n_{\max}(\boldsymbol{z}) \leq n_{\max}$ for all $\boldsymbol{z} \in \mathbb{S}$, where $n_{\max}$ may increase as a function of the number of nodes n provided $n_{\max} \leq n$. Then, for all $\delta > 0$, there exist $c > 0$ and $n_0 > 0$ such that, for all $n > n_0$,

$$\mathbb{P}\left(\max_{\boldsymbol{z}\in\mathbb{S}} |g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z}) - \mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \;\geq\; c\,(1+\delta)^{1/2}\,\|\mathcal{A}\|_\infty^2\, n_{\max}^2\, n\,(\log n)^{3/2}\right) \;\leq\; \epsilon,$$

where

$$\epsilon \;=\; 2\,\exp\left(-\delta\, n \log n\right).$$

Proof. To show that the probability mass of $g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})$ concentrates around its expectation $\mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})$, observe that the Lipschitz coefficient of the function $g : \mathbb{X} \times \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \mathbb{R}$ with respect to the Hamming metric $d : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}_0^+$ is given by

$$\|g\|_{Lip} \;=\; \sup_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\,\in\,\mathbb{X}\times\mathbb{X}:\; d(\boldsymbol{x}_1,\boldsymbol{x}_2)>0} \frac{|g(\boldsymbol{x}_1;\boldsymbol{\theta},\boldsymbol{z}) - g(\boldsymbol{x}_2;\boldsymbol{\theta},\boldsymbol{z})|}{d(\boldsymbol{x}_1,\boldsymbol{x}_2)}.$$

Since the term $\psi(\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})) - \psi(\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2{=}\mathbf{0},\boldsymbol{z}))$ of $g(\boldsymbol{x}_1;\boldsymbol{\theta},\boldsymbol{z})$ and $g(\boldsymbol{x}_2;\boldsymbol{\theta},\boldsymbol{z})$ cancels, we obtain

$$\frac{|g(\boldsymbol{x}_1;\boldsymbol{\theta},\boldsymbol{z}) - g(\boldsymbol{x}_2;\boldsymbol{\theta},\boldsymbol{z})|}{d(\boldsymbol{x}_1,\boldsymbol{x}_2)} \;=\; \frac{|\langle \boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}) - \boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2{=}\mathbf{0},\boldsymbol{z}), s(\boldsymbol{x}_1) - s(\boldsymbol{x}_2)\rangle|}{d(\boldsymbol{x}_1,\boldsymbol{x}_2)}.$$

By condition [C.1] and the fact that $\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}) - \boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2{=}\mathbf{0},\boldsymbol{z}) \in \mathbb{R}^{\dim(\boldsymbol{\eta})}$, there exists $c_0 > 0$ and $n_0 > 0$ such that, for all $n > n_0$,

$$\frac{|g(\boldsymbol{x}_1;\boldsymbol{\theta},\boldsymbol{z}) - g(\boldsymbol{x}_2;\boldsymbol{\theta},\boldsymbol{z})|}{d(\boldsymbol{x}_1,\boldsymbol{x}_2)} = \frac{|\langle \boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z}) - \boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2{=}\mathbf{0},\boldsymbol{z}), s(\boldsymbol{x}_1) - s(\boldsymbol{x}_2)\rangle|}{d(\boldsymbol{x}_1,\boldsymbol{x}_2)} \leq c_0\, n_{\max}(\boldsymbol{z}) \log n.$$

Therefore,

$$\|g\|_{Lip} \;\leq\; c_0\; n_{\max}(\boldsymbol{z})\; \log n \;\leq\; c_0\; n_{\max}\; \log n.$$

By construction of $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{x})$, the contributions of between-neighborhood subgraphs to the loglikelihood function are the same under both models, hence $g(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{z})$ reduces to a function of within-neighborhood edges which does not depend on between-neighborhood edges. Thus, by applying Lemma 3.1 to the Lipschitz function $g : \mathbb{X} \times \boldsymbol{\Theta} \times \mathbb{Z}$ of within-neighborhood edges with Lipschitz coefficient $\|g\|_{Lip} \;\leq\; c_0\;\; n_{\max}\, \log n$ with respect to the Hamming metric $d : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}_0^+$, there exists $c > 0$ such that, for all $t > 0$,

$$\mathbb{P}\left(|g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z}) - \mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \;\geq\; t\right) \;\leq\; 2\,\exp\left(-\frac{t^2}{c^2\; K\; n_{\max}^4\; \|\mathcal{A}\|_\infty^4\; (\log n)^2}\right).$$

A union bound over the $|\mathbb{S}| \leq K^n$ neighborhood structures shows that

$$\mathbb{P}\left(\max_{\boldsymbol{z}\in\mathbb{S}} |g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z}) - \mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \geq t\right) \;\leq\; 2\,\exp\left(-\frac{t^2}{c^2\,K\,n_{\max}^4\,\|\mathcal{A}\|_\infty^4\,(\log n)^2} + n\log K\right).$$

Choose $t = c\,(1+\delta)^{1/2}\,\|\mathcal{A}\|_\infty^2\,n_{\max}^2\,n\,(\log n)^{3/2}$, where $\delta > 0$. Then, for all $\delta > 0$,

$$\mathbb{P}\left(\max_{\boldsymbol{z}\in\mathbb{S}} |g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z}) - \mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \;\geq\; c\,(1+\delta)^{1/2}\,\|\mathcal{A}\|_\infty^2\,n_{\max}^2\,n\,(\log n)^{3/2}\right) \;\leq\; \epsilon,$$

where

$$\epsilon \;=\; 2\,\exp\left(-\delta\,n\log n\right).$$

Proposition 3.2 *Suppose that a random graph is governed by an exponential-family random graph model with countable support $\mathbb{X}$ and local dependence satisfying conditions [C.1] and [C.2]. Then there exist $c > 0$ and $n_0 > 0$ such that, for all $n > n_0$,*

$$\max_{\boldsymbol{z}\in\mathbb{S}} |\mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \;\leq\; c\,K\,n_{\max}^2\,\log n.$$

Proof. By definition,

$$\mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z}) \;=\; \mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X}).$$

By construction of $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{x})$ and $p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{x})$, the contributions of between-neighborhood subgraphs to the loglikelihood function are the same under both models, hence the expectation of the loglikelihood ratio reduces to the expectation of the loglikelihood ratio of within-neighborhood subgraphs:

$$\mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X}) = \sum_{k=1}^{K}\left[\mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}_{k,k}) - \mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X}_{k,k})\right].$$

By the triangle inequality,

$$\left|\mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}) - \mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X})\right| \leq \sum_{k=1}^{K}\left|\mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}_{k,k}) - \mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\boldsymbol{0},\boldsymbol{z})}(\boldsymbol{X}_{k,k})\right|.$$

The terms $|\mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}_{k,k}) - \mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{X}_{k,k})|$ can be bounded above as follows:

$$\begin{aligned}\left|\mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}_{k,k}) - \mathbb{E}\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2=\mathbf{0},\boldsymbol{z})}(\boldsymbol{X}_{k,k})\right| &\leq |\langle \boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k},\boldsymbol{z}) - \boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k,0},\boldsymbol{z}),\, \mathbb{E}\, s_{k,k}(\boldsymbol{X})\rangle| \\ &+ |\psi_{k,k}(\boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k},\boldsymbol{z})) - \psi_{k,k}(\boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k,0},\boldsymbol{z}))|,\end{aligned}$$

where $\boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k},\boldsymbol{z})$, $s_{k,k}(\boldsymbol{x})$, and $\psi_{k,k}(\boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k},\boldsymbol{z}))$ are the natural parameter vector, the sufficient statistics vector, and the log-normalizing constant of $p_{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{z})}(\boldsymbol{X}_{k,k})$, and $\boldsymbol{\theta}_{k,k,0} = (\theta_{1,k,k}, \boldsymbol{\theta}_{2,k,k} = \mathbf{0})$. We bound the two terms on the right-hand side of the inequality above one by one.

First term. By condition [C.2], there exist $c_1 > 0$, $c_2 > 0$, and $n_1 > 0$ such that, for all $n > n_1$,

$$\begin{aligned}|\langle \boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k},\boldsymbol{z}) - \boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k,0},\boldsymbol{z}),\, \mathbb{E}\, s_{k,k}(\boldsymbol{X})\rangle| &\leq c_1\, \|\boldsymbol{\theta}_{k,k} - \boldsymbol{\theta}_{k,k,0}\|_2\, n_{\max}(\boldsymbol{z})^2\, \log n \\ &\leq c_2\, n_{\max}^2\, \log n,\end{aligned}$$

where the last inequality follows from the assumption that $\boldsymbol{\Theta}_{k,k}$ is compact.

Second term. By the mean-value theorem along with classic exponential-family properties, there exists $\dot{\boldsymbol{\eta}}_{k,k} = \alpha\, \boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k},\boldsymbol{z}) + (1-\alpha)\, \boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k,0},\boldsymbol{z})$ $(0 < \alpha < 1)$ such that

$$|\psi_{k,k}(\boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k},\boldsymbol{z})) - \psi_{k,k}(\boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k,0},\boldsymbol{z}))| = |\langle \boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k},\boldsymbol{z}) - \boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k,0},\boldsymbol{z}),\, \mathbb{E}_{\dot{\boldsymbol{\eta}}_{k,k}}\, s_{k,k}(\boldsymbol{X})\rangle|.$$

Therefore, the second term can be bounded along the same lines as the first term, which implies that there exist $c_3 > 0$ and $n_2 > 0$ such that, for all $n > n_2$,

$$|\psi_{k,k}(\boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k},\boldsymbol{z})) - \psi_{k,k}(\boldsymbol{\eta}_{k,k}(\boldsymbol{\theta}_{k,k,0},\boldsymbol{z}))| \leq c_3\, n_{\max}^2\, \log n.$$

Conclusion. Collecting terms shows that that there exist $c > 0$ and $n_0 = \max(n_1, n_2) > 0$ such that, for all $n > n_0$,

$$\max_{\boldsymbol{z}\in\mathbb{S}} |\mathbb{E}\; g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \leq c\, K\, n_{\max}^2\, \log n.$$

Armed with Propositions 3.1 and 3.2, we can prove Theorem 3.1.

PROOF OF THEOREM 3.1.

Observe that, for all $t > 0$,

$$\begin{aligned}
& \mathbb{P}\left(\max_{\boldsymbol{z}\in\mathbb{S}} |g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \;\geq\; t\right) \\
\leq\;& \mathbb{P}\left(\max_{\boldsymbol{z}\in\mathbb{S}} |g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z}) - \mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| + \max_{\boldsymbol{z}\in\mathbb{S}} |\mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \;\geq\; t\right) \\
\leq\;& \mathbb{P}\left(\max_{\boldsymbol{z}\in\mathbb{S}} |g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z}) - \mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \;\geq\; \frac{t}{2}\right) + \mathbb{P}\left(\max_{\boldsymbol{z}\in\mathbb{S}} |\mathbb{E}\, g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \;\geq\; \frac{t}{2}\right).
\end{aligned}$$

Choose $t = c\ (1+\delta)^{1/2}\ \|\mathcal{A}\|_\infty^2\ n_{\max}^2 n\ (\log n)^{3/2}$, where $c > 0$ is identical to the constant c in Proposition 3.1 and $\delta > 0$. Then, by Propositions 3.1 and 3.2, for all $\delta > 0$, there exists $n_0 > 0$ such that, for all $n > n_0$,

$$\mathbb{P}\left(\max_{\boldsymbol{z}\in\mathbb{S}} |g(\boldsymbol{X};\boldsymbol{\theta},\boldsymbol{z})| \;\geq\; c\,(1+\delta)^{1/2}\,\|\mathcal{A}\|_\infty^2\, n_{\max}^2\, n\,(\log n)^{3/2}\right) \;\leq\; 2\,\exp\left(-\frac{\delta\, n \log n}{4}\right).$$

Chapter 4

Discussion: directions for future research

In the previous two chapters, novel methods and theory were introduced for two classes of models of high-dimensional and dependent data: vector autoregressive processes and exponential-family random graph models. Both models were endowed with additional structure for the purpose of constructing scalable methods with desirable statistical properties. In both cases, simulation studies and applications to large data sets were presented in order to demonstrate that the proposed model estimation approaches both work fast and lead to better models compared to the existing methods. However, many open problems remain for both models. An overview of some important directions for future research of vector autoregressive processes and exponential-family random graph models is given in Sections 4.1 and 4.2 respectively.

4.1 Directions for future research of high-dimensional multivariate time series

An important direction for future research of vector autoregressive processes with additional structure is to investigate the impact of dependence on stability selection (Meinshausen and Bühlmann, 2010). While the theoretical results in Section 2.4 show that the scaling of the regularization parameters of the two-step ℓ_1-penalized least squares method depends on the unknown values of $\boldsymbol{\beta}^\star$ and $\boldsymbol{\Sigma}$, in practice, the selection of regularization parameters is either performed by cross-validation or sidestepped by using stability selection.

Simulation studies in Chapter 2 identified that stability selection works well when the order of vector autoregressive processes is 1 and the sparsity is between 1% and 2% (see, e.g., Table 2.2 in Section 2.5). However, when the order of vector autoregressive processes is larger and the

vector autoregressive processes are sparser, the dependence might impact stability selection and increase the number of false-positive edges in the first step of the two-step ℓ_1-penalized least squares method, which in turn might give rise to overestimates of the radius of dependence. It would be interesting to explore approaches to stability selection that can capture the dependence induced by vector autoregressive processes. One approach is to divide the time-dependent observations into blocks that capture the dependence of the observations, as suggested by Künsch (1989) and Meinshausen and Bühlmann (2010, p. 471). Some insights on how to construct such blocks of time-dependent observations can be found in, e.g., Künsch (1989), Politis and Romano (1994), and Davis et al. (2012).

Another interesting direction for future research is to explore promising extensions of the proposed two-step ℓ_1-penalized least squares method. One interesting extension would be to impose a parametric form on the transition matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L$ and the variance-covariance matrix $\boldsymbol{\Sigma}$, i.e., to allow $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L$ and $\boldsymbol{\Sigma}$ to depend on distance in some parametric form. To do so would require additional model assumptions, but it could reduce statistical error.

A second interesting extension would be to assume that the radius of the past-present dependence captured by $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L$ may not be the same as the radius of the present-present dependence captured by $\boldsymbol{\Sigma}$. Such extensions would make sense in applications where the present-present dependence captured by $\boldsymbol{\Sigma}$ is more local than the past-present dependence captured by $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L$.

A third and most ambitious extension would be to go beyond vector autoregressive processes and to extend the two-step ℓ_1-penalized least squares method to other high-dimensional models, such as high-dimensional regression models and high-dimensional graphical models (e.g., Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010). That can be performed as long as additional structure of the form considered in Chapter 2 is available and consistent model selection in high dimensions is possible.

4.2 Directions for future research of exponential-family random graph models

The two-step likelihood-based approach proposed in Chapter 3 enables massive-scale estimation of exponential-family random graph models with unknown neighborhood structure provided that the number of neighborhoods K is known. An important goal of future research is to develop methods for selecting K when K is unknown. Even in the special case of stochastic block models, the issue of selecting K has not received much attention—with the exception of recent work by Saldana et al. (2017) and Wang and Bickel (2017). Extending such methods to the more general models considered in Chapter 3 would be useful.

In addition, an important direction for future research is to relax the assumption of non-overlapping neighborhoods introduced in Chapter 3. While it makes sense for many applications, some networks do not admit a restrictive division of the set of nodes into disjoint communities. One telling example is an academic collaboration network: many researchers do not belong to just a single field of study, since their research contributes to multiple disciplines. Thus, collaborations of those researchers with scientists in one field are not necessarily independent of collaborations with scientists working in a different field.

Last, but not least, one of the major issues of modeling large networks is that model terms suitable for small networks might not be flexible enough to represent dependencies encountered in large networks. Thus, the problem of selecting meaningful specifications for exponential-family random graph models with local dependence has to be addressed in order to perform successful analyses of large real-world data sets.

Indeed, in the case of large networks, a good model has to successfully capture transitivity, degree heterogeneity, homophily, and other distinguishing attributes of real-world networks. Multiple attempts were made to introduce specifications of exponential-family random graph models, which could express those unique features in a way amenable for statistical inference. The most notable works in this direction include Snijders et al. (2006), Hunter and Handcock (2006), and Snijders et al. (2010). However, historically most of the studies and practical applications of exponential-

family random graph models focused on small or sparse networks (e.g., the large social network considered in Goodreau (2007) has 1,681 nodes and only 1,236 edges). Therefore, the optimal choice of suitable model terms and its effect on the performance of exponential-family random graph models in the case of large and non-sparse graphs are not well understood.

Nevertheless, the exponential-family random graph framework permits a great variety of model terms, some combination of which might be capable of modeling large networks. An important direction for future research is to identify which classes of exponential-family random graph models are sufficiently flexible to handle large networks. In particular, it might be beneficial to perform a study of the joint behavior of terms modeling transitive closure, degree heterogeneity, and homophily in the case of large networks. It would also be of interest to assess the effect of increasing edge density on the performance of various model terms. For example, in denser networks, the number of edges and transitive edges may be very similar or even equal. This is problematic on both computational and theoretical grounds and can result in the non-existence of estimators or estimators, which are computationally difficult to obtain.

Chapter 5

Supplementary materials

The data along with all R source code used in Chapter 2 is contained in the supplementary archive available online at

http://amstat.tandfonline.com/doi/suppl/10.1080/10618600.2016.1265528/suppl_file/ucgs_a_1265528_sm0228.zip

The R source code requires R (R Core Team, 2017) packages mAr (Barbosa, 2012), Matrix (Bates and Maechler, 2016), lars (Hastie and Efron, 2013), spcov (Bien and Tibshirani, 2012), glmnet (Friedman et al., 2010), tsDyn (Stigler, 2010), expm (Goulet et al., 2015), and parallel (R Core Team, 2017). The application requires in addition the R packages rworldmap (South, 2011) and timeSeries (Rmetrics Core Team et al., 2015).

The data and R source code used in Chapter 3 is contained in the supplementary archive available online at

https://github.com/kasht/ERGM-supplement.

Note that the archive contains an updated version of R package hergm (Schweinberger and Luna, 2015). The basic version of this package can be found on the CRAN website. The R source code additionally requires R packages car (Fox and Weisberg, 2011) and parallel (R Core Team, 2017).

Bibliography

American Lung Association (2015), "State of the air: 2015," Tech. rep., American Lung Association.

Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013), "Pseudo-likelihood methods for community detection in large sparse networks," *The Annals of Statistics*, 41, 2097–2122.

Babkin, S., and Schweinberger, M. (2017), "Massive-scale estimation of exponential-family random graph models with local dependence," `http://arxiv.org/abs/1703.09301`.

Bańbura, M., Giannone, D., and Reichlin, L. (2010), "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25, 71–92.

Barbosa, S. M. (2012), *Mar: Multivariate Autoregressive Analysis*, R package version 1.1-2.

Basu, S., and Michailidis, G. (2015), "Regularized estimation in sparse high-dimensional time series models," *The Annals of Statistics*, 43, 1535–1567.

Bates, D., and Maechler, M. (2016), *Matrix: Sparse and Dense Matrix Classes and Methods*, R package version 1.2-4.

Bickel, P. J., and Chen, A. (2009), "A nonparametric view of network models and Newman-Girvan and other modularities," *Proceedings of the National Academy of Sciences*, 106, 21068–21073.

Bickel, P. J., Choi, D., Chang, X., and Zhang, H. (2013), "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels," *The Annals of Statistics*, 41, 1922–1943.

Bien, J., and Tibshirani, R. (2012), *Spcov: Sparse Estimation of a Covariance Matrix*, R package version 1.01.

Bollobás, B. (1998), *Modern Graph Theory*, Springer.

Brillinger, D. R. (2001), *Time Series: Data Analysis and Theory*, Philadelphia: Society for Industrial and Applied Mathematics.

Brown, L. (1986), *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*, Hayworth, CA, USA: Institute of Mathematical Statistics.

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, New York: Springer.

Celisse, A., Daudin, J. J., and Pierre, L. (2012), "Consistency of maximum-likelihood and variational estimators in the stochastic block model," *Electronic Journal of Statistics*, 6, 1847–1899.

Chatterjee, S., and Diaconis, P. (2013), "Estimating and understanding exponential random graph models," *The Annals of Statistics*, 41, 2428–2461.

Chen, G., Wan, X., Yang, G., and Zou, X. (2015), "Traffic-related air pollution and lung cancer: a meta-analysis," *Thoracic Cancer*, 6, 307–318.

Daudin, J. J., Picard, F., and Robin, S. (2008), "A mixture model for random graphs," *Statistics and Computing*, 18, 173–183.

Davis, R. A., Mikosch, T., and Cribben, I. (2012), "Towards estimating extremal serial dependence via the bootstrapped extremogram," *Journal of Econometrics*, 170, 142–152.

Davis, R. A., Zang, P., and Zheng, T. (2016), "Sparse vector autoregressive modeling," *Journal of Computational and Graphical Statistics*, 25, 1077–1096.

De Mol, C., Giannone, D., and Reichlin, L. (2008), "Forecasting using a large number of predictors: is Bayesian shrinkage a valid alternative to principal components?" *Journal of Econometrics*, 146, 318–328.

Eichler, M. (2012), "Graphical modelling of multivariate time series," *Probability Theory and Related Fields*, 153, 233–268.

Ensor, K. B., Raun, L., and Persse, D. (2013), "A case-crossover analysis of out-of-hospital cardiac arrest and air pollution," *Circulation*, 1192–1199.

Fox, J., and Weisberg, S. (2011), *An R Companion to Applied Regression*, Thousand Oaks CA: Sage, 2nd ed.

Frank, O., and Strauss, D. (1986), "Markov graphs," *Journal of the American Statistical Association*, 81, 832–842.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 33, 1–22.

Friston, K. (2009), "Causal modelling and brain connectivity in functional magnetic resonance imaging," *PLOS Biology*, 7, 220–225.

Goodreau, S. M. (2007), "Advances in exponential random graph models applied to a large social network," *Social Networks*, 29, 231–248.

Goulet, V., Dutang, C., Maechler, M., Firth, D., Shapira, M., and Stadelmann, M. (2015), *Expm: Matrix Exponential*, R package version 0.999-0.

Granovetter, M. (1973), "The strength of weak ties," *American Journal of Sociology*, 78, 1360–1380.

Häggström, O., and Jonasson, J. (1999), "Phase transition in the random triangle model," *Journal of Applied Probability*, 36, 1101–1115.

Handcock, M. (2003), "Assessing degeneracy in statistical models of social networks," Tech. rep., Center for Statistics and the Social Sciences, University of Washington, http://www.csss.washington.edu/Papers.

Hastie, T., and Efron, B. (2013), *Lars: Least Angle Regression, Lasso and Forward Stagewise*, R package version 1.2.

Hoek, G., Krishnan, R., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., and Kaufman, J. (2013), "Long-term air pollution exposure and cardio-respiratory mortality: a review," *Environmental Health*, 12, 1–15.

Holland, P. W., and Leinhardt, S. (1981), "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical Association*, 76, 33–65.

Hunter, D. R., and Handcock, M. S. (2006), "Inference in curved exponential family models for networks," *Journal of Computational and Graphical Statistics*, 15, 565–583.

Hunter, D. R., and Lange, K. (2004), "A tutorial on MM algorithms," *The American Statistician*, 58, 30–38.

Jonasson, J. (1999), "The random triangle model," *Journal of Applied Probability*, 36, 852–876.

Kolaczyk, E. D. (2009), *Statistical Analysis of Network Data: Methods and Models*, New York: Springer.

Koop, G. M. (2013), "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics*, 28, 177–203.

Koschade, S. (2006), "A social network analysis of Jemaah Islamiyah: the applications to counterterrorism and intelligence," *Studies in Conflict & Terrorism*, 29, 559–575.

Künsch, H. R. (1989), "The jackknife and the bootstrap for general stationary observations," *The Annals of Statistics*, 17, 1217–1241.

Lei, J., and Rinaldo, A. (2015), "Consistency of spectral clustering in stochastic block models," *The Annals of Statistics*, 43, 215–237.

Liang, F., Jin, I. H., Song, Q., and Liu, J. S. (2016), "An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants," *Journal of the American Statistical Association*, 111, 377–393.

Loh, P. L., and Wainwright, M. J. (2012), "High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity," *The Annals of Statistics*, 40, 1637–1664.

Lusher, D., Koskinen, J., and Robins, G. (2013), *Exponential Random Graph Models for Social Networks*, Cambridge, UK: Cambridge University Press.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer Science & Business Media.

— (2011), *Vector Autoregressive Models*, Springer.

Meinshausen, N., and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the LASSO," *The Annals of Statistics*, 34, 1436–1462.

— (2010), "Stability selection," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 72, 417–473.

Negahban, S., and Wainwright, M. J. (2011), "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, 39, 1069–1097.

Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2014), "Spatio-temporal data fusion for very large remote sensing datasets," *Technometrics*, 56, 174–185.

Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087.

Pattison, P., and Robins, G. (2002), "Neighborhood-based models for social networks," *Sociological Methodology*, 32, 301–337.

Politis, D. N., and Romano, J. P. (1994), "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rao, S. T., Zurbenko, I. G., Neagu, R., Porter, P. S., Ku, J. Y., and Henry, R. F. (1997), "Space and time scales for ambient ozone data," *Bulletin of the American Meteorological Society*, 78, 2153–2166.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. (2010), "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *The Annals of Statistics*, 38, 1287–1319.

Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009), "On the geometry of discrete exponential families with application to exponential random graph models," *Electronic Journal of Statistics*, 3, 446–484.

Rmetrics Core Team, Wuertz, D., Setz, T., and Chalabi, Y. (2015), *Timeseries: Rmetrics - Financial Time Series Objects*, R package version 3022.101.2.

Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral clustering and the high-dimensional stochastic block model," *The Annals of Statistics*, 39, 1878–1915.

Saldana, D. F., Yu, Y., and Feng, Y. (2017), "How many communities are there?" *Journal of Computational and Graphical Statistics*, accepted.

Schweinberger, M. (2011), "Instability, sensitivity, and degeneracy of discrete exponential families," *Journal of the American Statistical Association*, 106, 1361–1370.

— (2017), "Consistent structure estimation of exponential-family random graph models with additional structure," Tech. rep., Department of Statistics, Rice University.

Schweinberger, M., Babkin, S., and Ensor, K. B. (2017), "High-dimensional multivariate time series with additional structure," *Journal of Computational and Graphical Statistics*, accepted.

Schweinberger, M., and Handcock, M. S. (2015), "Local dependence in random graph models: characterization, properties and statistical inference," *Journal of the Royal Statistical Society B*, 77, 647–676.

Schweinberger, M., and Luna, P. (2015), "HERGM: Hierarchical exponential-family random graph models," Tech. rep., Department of Statistics, Rice University.

Schweinberger, M., and Stewart, J. (2016), "Consistent M-estimation of curved exponential-family random graph models with local dependence and growing neighborhoods," `http://arxiv.org/abs/1702.01812`.

Shalizi, C. R., and Rinaldo, A. (2013), "Consistency under sampling of exponential random graph models," *The Annals of Statistics*, 41, 508–535.

Sims, C. A. (1980), "Macroeconomics and reality," *Econometrica*, 1–48.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New specifications for exponential random graph models," *Sociological Methodology*, 36, 99–153.

Snijders, T. A. B., van de Bunt, G., and Steglich, C. E. G. (2010), "Introduction to stochastic actor-based models for network dynamics," *Social Networks*, 32, 44–60.

Song, S., and Bickel, P. J. (2011), "Large vector auto regressions," `http://arxiv.org/abs/1106.3915`.

South, A. (2011), "Rworldmap: a new R package for mapping global data," *The R Journal*, 3, 35–43.

Stigler, M. (2010), *Tsdyn: Threshold Cointegration: Overview and Implementation in R*, R package version 0.7-2.

Stock, J. H., and Watson, M. W. (2006), "Forecasting with many predictors," *Handbook of Economic Forecasting*, 1, 515–554.

Strauss, D. (1986), "On a general class of models for interaction," *SIAM Review*, 28, 513–527.

Thiemichen, S., and Kauermann, G. (2017), "Stable exponential random graph models with non-parametric components for large dense networks," *Social Networks*, 49, 67–80.

Thompson, S. K. (2012), *Sampling*, Wiley, 3rd ed.

Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2013), "Model-based clustering of large networks," *The Annals of Applied Statistics*, 7, 1010–1039.

Waggoner, D. F., and Zha, T. (1999), "Conditional forecasts in dynamic multivariate models," *Review of Economics and Statistics*, 81, 639–651.

Wang, Y. X., and Bickel, P. J. (2017), "Likelihood-based model selection for stochastic block models," *The Annals of Statistics*, accepted.

Wasserman, S., and Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press.

Watson, M. W. (1994), "Vector autoregressions and cointegration," *Handbook of Econometrics*, 4, 2843–2915.

Wilson, G. T., Reale, M., and Haywood, J. (2015), *Models for Dependent Time Series*, CRC Press.

World Health Organization (2014), "7 million premature deaths annually linked to air pollution," Tech. rep., World Health Organization.

Yang, J., and Leskovec, J. (2015), "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, 42, 181–213.